

Résumé

La prise en compte des besoins, des intentions et des spécificités cognitives, culturelles ou autres qui caractérisent le profil d'un utilisateur constitue un élément déterminant pour améliorer la pertinence des réponses lors d'une session de Recherche d'Information dans des grandes bases de documents. La modélisation des profils utilisateurs et la manière de les adapter à différents utilisateurs qui n'ont pas une idée précise sur l'information qu'ils recherchent, nous permet d'offrir un accès personnalisé au contenu de documents scientifiques fondé sur l'exploitation du profil utilisateur. Notre propos, dans ce mémoire, est de proposer un modèle de l'utilisateur fondé sur les connaissances et un système implémentant le Raisonnement à Partir des Cas pour acquérir ces connaissances, les structurer et les faire évoluer.

Mots-clés : Profil utilisateur, Modèle utilisateur, Recherche d'information, RaPC, Base de cas, Personnalisation de l'information

Abstract

Taking into account the needs, the intention and cognitive, cultural or various specificities to characterize the user profile, is a major challenge to improve the relevance of information retrieval systems. The users' models and the way to adapt them to different users (who need help to build their request during query processing) allow using a personalised access to scientific documents based on user profile. So, we propose a user model based on users' knowledge and users' preferences. We have defined a system based on CBR to capture users' preferences and knowledge, to structure them and to manage the user profile evolution. We validate some results by means of a prototype.

Keywords: User profile, User model, CBR, Information retrieval, Personalization of information.

Cette recherche a été partiellement soutenue par le Ministère délégué à la Recherche et aux Nouvelles Technologies, dans le programme ACI Masses de Données, projet #MD-33.

Table des matières

Chapitre 1 : Présentation générale	4
1.1 Introduction.....	4
1.2. Contexte du travail.....	4
Chapitre 2 : Modélisation et évolution des profils.....	5
2.1 Notion de profil utilisateur.....	5
2.2 Typologie des connaissances constituant le profil.....	6
2.3. Modèles formels de représentation d'un profil.....	8
2.4. Méthodes de définition et d'évolution d'un profil.....	9
2.5. Historique des systèmes de modélisation de l'utilisateur.....	15
Chapitre 3 : La modélisation de l'utilisateur – Notre proposition.....	17
3.1. Typologie de connaissances impliquées	17
3.2. Formalisation et construction du modèle de l'utilisateur ainsi qu'évolution d'un profil	22
Chapitre 4. Evaluation.....	28
Chapitre 5 : Conclusion et perspectives.....	30
Bibliographie.....	31
Annexe. Listes.....	32

Chapitre 1 : Présentation générale

1.1 Introduction

L'accès à une information pertinente adaptée aux besoins et au contexte de l'utilisateur est un défi dans un environnement Internet caractérisé par une prolifération de ressources hétérogènes conduisant à des volumes considérables. Cette masse de données accrue et les divers types de données impliquent des réponses de plus en plus lentes et pas toujours pertinentes de la part des systèmes de recherche d'information. Plusieurs groupes de recherche travaillent pour palier ces problèmes. Par exemple le CNRS a mené deux actions spécifiques concernant la personnalisation de l'information (AS98) et le passage à l'échelle dans les systèmes de recherche d'information (AS91). Les deux groupes de travail sont arrivés à la même conclusion : une des principales pistes de recherches favorisant le passage à l'échelle des systèmes de recherche d'information est l'exploitation généralisée de préférences définies dans les profils des utilisateurs. Dans ce cadre le Ministère Délégué à la Recherche et aux Nouvelles Technologies a organisé la cellule ACI Masses de Données, qui comprend entre autres le projet « Accès personnalisé à des masses de données » (APMD). C'est dans le cadre de ce projet que se situe le travail présenté dans ce mémoire.

1.2. Contexte du travail

Le projet APMD (<http://apmd.prism.uvsq.fr>) a démarré en novembre 2004 et a pour objectif de promouvoir l'usage des profils des utilisateurs dans l'évaluation des requêtes en environnement de masses de données. Il s'articule autour de trois axes : (i) une exploration qui vise la construction et la structuration des concepts de profil d'utilisateur et de la qualité de l'information, (ii) l'exploitation de ces concepts dans le cycle de vie d'une requête et (iii) la validation des propositions des partenaires. Effectivement ce projet rassemble des partenaires provenant des communautés de la Recherche de l'Information et des bases de données. Le LIRIS est un des partenaires. Notre travail dans ce laboratoire s'inscrit parfaitement dans le cadre du projet APMD en suivant ses objectifs et ses axes de travail.

APMD identifie quatre sous-projets correspondant à quatre domaines : (i) la modélisation et l'évolution des profils, (ii) l'exécution adaptative des requêtes tenant compte des profils et du passage à l'échelle, (iii) la prise en compte de la qualité dans la personnalisation, ainsi que (iiii) l'évaluation et la validation. Nous sommes concernés par les sous-projets 1, 2 et 4 qui vont être développés par la suite selon notre point de vue.

L'objectif du sous-projet 1 est l'étude de la notion du profil qui constitue la base de la personnalisation de l'information dans un domaine d'application. Notre domaine est celui des bibliothèques numériques, et particulièrement celle de DOC'INSA, dans le cadre du projet CITHER (<http://csidoc.insa-lyon.fr/these/>). Le sous-projet 2 vise à faciliter l'écriture des requêtes de l'utilisateur en les enrichissant avec des données invariantes, communes à toutes les requêtes, mais spécifiques à un utilisateur, ou un groupe d'utilisateurs. C'est là que nous allons mettre en évidence les avantages du Raisonnement à Partir des Cas et de l'utilisation des stéréotypes d'utilisateurs pour la recherche de l'information.

Le sous-projet 4 a pour objectif d'évaluer et de valider notre approche par la réalisation d'un prototype et la définition des scénarios d'évaluation. Nos tests s'effectueront à partir du corpus de thèses numériques de CITHER.

Chapitre 2 : Modélisation et évolution des profils

Ce chapitre comporte une étude (i) de la typologie des connaissances constituant un profil, (ii) des modèles formels de représentation d'un profil, et (iii) des méthodes de définition et d'évolution d'un profil. Mais avant tout nous allons tenter de définir la notion du profil et de sa place dans la Recherche de l'Information.

2.1 Notion de profil utilisateur

Quand les utilisateurs se connectent à un système de recherche d'information, ils apportent leurs bagages intellectuels par rapport à un sujet précis d'investigation, ainsi que leurs préférences de pertinences et de qualité. Toutes ces variations qui caractérisent un utilisateur ou un groupe d'utilisateurs, peuvent se regrouper sous le terme de profil de l'utilisateur.

Mais l'utilisateur peut collecter l'information de différentes manières : par l'utilisation des mots-clés dans les moteurs de recherche, par la réception des documents par e-mail, ou encore par navigation exploratoire sur Internet. Ce sont des moyens plutôt volontaires pour accéder à l'information. Provoquer la rencontre entre l'utilisateur et l'information pertinente d'une manière involontaire est plutôt l'axe de recherche des systèmes de filtrage. Selon Maes¹, « le filtrage est un processus qui consiste à extraire les informations pertinentes et de qualité à partir d'une imposante masse d'informations ». Le filtrage et la recherche de l'information sont très proches. Ils ont chacun leurs spécificités résumées dans le **Tableau 1** [GAU et STE 2003].

	<i>Recherche de l'information</i>	<i>Filtrage basé sur le contenu</i>
Approche	Trouver l'information recherchée	Filtrer l'information non désirée
Livraison	Corpus statique sur demande	Flux dynamique
Persistance	Des besoins à court terme	Des intérêts à long terme
Personnalisation	Non personnalisé	Profil d'utilisateur requis
Analyse de contenu	Utilise souvent des mots-clés	Différents et multiples dispositifs utilisés p.ex. l'évaluation et le retour de pertinence
Fonctionnalités	Non personnalisé Non adaptatif Non dynamique A court terme	Personnalisé S'adapte aux changements du profil utilisateur Filtre dynamiquement l'information entrante A long terme

Tableau 1 . Comparaison des principes de recherche d'information et de filtrage de l'information fondés sur le contenu

Comme indiqué dans ce tableau, le profil de l'utilisateur et ses intérêts prennent une place prépondérante dans les systèmes de filtrage de l'information. Grâce à la modélisation de l'utilisateur, le système doit être capable de sélectionner les informations à transmettre et d'adapter

¹ <http://www.loria.fr/equipes/maia/lexique/fullpage.php>

la restitution. On peut ainsi envisager la notion de personnalisation de l'information comme un processus de définition, de construction et d'utilisation des profils, pour répondre de façon efficace et adaptée à un même type de requête émise par des utilisateurs de profils différents. C'est pourquoi, nous allons traiter le profil de l'utilisateur dans ce contexte.

Les liens explicites entre le filtrage et l'utilisateur apparaissent dans la figure suivante (Figure 1) <http://www.loria.fr/equipes/maia/lexique/fullpage.php>.

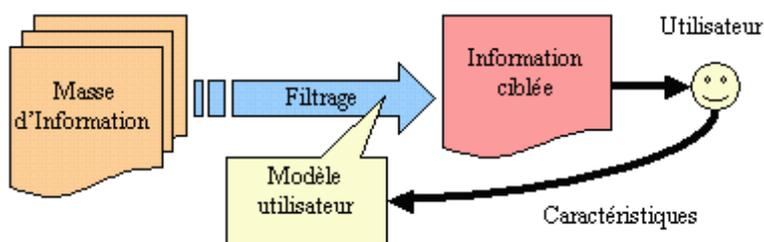


Figure 1 . Liens entre le filtrage et le modèle de l'utilisateur

Le profil peut être constitué d'un ensemble de caractéristiques avec des valeurs associées contenant ce que l'utilisateur préfère, ce qu'il est capable de faire et de comprendre, et d'autres informations personnelles le concernant. Les caractéristiques de l'utilisateur peuvent être obtenues de différentes façons en fonction de l'autonomie du système, et de ses capacités d'observation et d'adaptation.

Ainsi nous pouvons distinguer les profils statiques qui sont renseignés par un expert avant la mise en exploitation du système, les profils réflexifs renseignés par les utilisateurs eux-mêmes par le biais de formulaires, et les profils corrigés dynamiquement selon lequel un sous-système de modélisation observe l'utilisateur de derrière l'interface et apprend son profil à partir de ses actions. L'approche statique ne permet pas la mise à jour des profils utilisateurs après le démarrage du système. La deuxième approche permet plus de précision et d'adaptation alors que la troisième s'inscrit mieux dans le cadre des systèmes de filtrage favorisant l'accès à l'information d'une façon spontanée. Notre but est de construire un système qui pourrait prendre en compte ces trois approches de construction du profil de l'utilisateur.

La modélisation d'un utilisateur d'un système d'information a fait l'objet de nombreux travaux [FINK00] [FINK02] [KOB01] [KOB03]. C'est une problématique complexe nécessitant les contributions de différents types de connaissances apportées par de multiples branches de sciences de l'information et de l'homme.

2.2 Typologie des connaissances constituant le profil

Les connaissances utilisées pour la modélisation de l'utilisateur proviennent d'une part des sciences humaines, mais aussi des disciplines scientifiques plus techniques. A partir de la littérature, 3 types de connaissances ont attiré notre attention. Elles proviennent du monde de la sociologie, de la psychologie cognitive et de l'Intelligence Artificielle. Elles sont fondées sur différents points de vue, mais elles ne sont pas complètement disjointes et l'utilisateur peut conjuguer les apports de ces différentes approches dans ses activités de recherche.

2.2.1. Apports de la sociologie

Les sociologues fédérés par Carberry [CAR88] proposent un modèle d'actions ou de tâches de l'utilisateur, en construisant une hiérarchie de **stéréotypes** de l'utilisateur. Pour chaque stéréotype ou groupe d'utilisateurs, ils associent un plan d'actions possibles.

Ensuite KOBSA [KOB89] définit trois étapes pour la conception d'un stéréotype :

- Identifier des groupes d'utilisateurs et situer chaque utilisateur dans un groupe.
- Identifier des caractéristiques clés de chaque population ou groupe.

- La collection des caractéristiques d'une population d'utilisateur est appelée « stéréotype ». Ces stéréotypes sont alors ordonnés en une hiérarchie.

Donc, un stéréotype est une représentation d'un groupe de personnes en tant qu'individu unique qui n'existe pas comme personne réelle, mais seulement comme une extraction de la plupart des caractéristiques communes d'un groupe d'utilisateurs.

Kobsa propose également des méthodes pour collecter les informations sur le comportement des utilisateurs qui sont :

- La bibliothèque des plans : ce sont des plans préétablis dans le système. Les séquences observées des actions des utilisateurs sont comparées avec ces plans.

- La construction des plans : en plus d'une bibliothèque de tous les plans possibles avec les effets et les pré-conditions des actions, le système peut intégrer toute séquence d'actions accomplie par l'utilisateur.

KOBSA et *CARBBERY* mettent en évidence le fait que le système peut identifier le stéréotype de l'utilisateur au moment de l'interaction, et que les stéréotypes pourront être rassemblés par le système de session en session et ainsi contribuer à la construction d'un modèle final.

Eliane RICH [RICH1979], [RICH1983], [RICH1989] a également fait un travail remarquable sur la définition et l'utilisation des stéréotypes pour la modélisation de l'utilisateur. Elle souligne qu'un stéréotype est un trait (ou une caractéristique commune) partagé par plusieurs utilisateurs [RICH1979]. Un stéréotype utilisateur implique pour le système de connaître deux types d'information : les propriétés du stéréotype à capter, ainsi que l'événement ou le comportement qui implique un stéréotype particulier. *Si cette information est hautement dynamique et dépendante du domaine, une approche par groupement est préférable à une approche statistique, dès qu'il est possible d'identifier automatiquement les catégories relatives et d'adapter à une autre population d'utilisateurs, leurs préférences et leurs caractéristiques.*

La communauté de la modélisation de l'utilisateur a souvent réutilisé cette notion de stéréotypes développés par RICH [KOB01], [FINK00] L'utilisation des stéréotypes est aussi communément utilisée dans le filtrage de l'information sur le web, comme un moyen de classifier les utilisateurs. Ainsi, parmi les utilisateurs de DOC'INSA nous pouvons identifier clairement des stéréotypes associés à diverses catégories d'utilisateurs comme les professeurs, les doctorants, les visiteurs etc... C'est pourquoi, l'utilisation des stéréotypes sera un point important dans notre approche de construction de profil.

2.2.2. Apport des sciences cognitives

L'approche cognitive permet d'envisager un modèle des connaissances de l'utilisateur, en définissant des corrélations entre les différents types de connaissances exploitées par l'utilisateur et le comportement de ce dernier face à une situation de recherche documentaire. Les études cognitives en sciences de l'information remontent à 1977 avec *BELLKIN* dont l'un des objectifs majeurs était d'apprendre le processus informationnel individuel et par la suite de l'illustrer par un modèle [BEL 87].

Bien plus tard, Allen distingue quatre catégories de connaissances [ALL 91] :

- Les connaissances de l'utilisateur sur son propre monde qui ont suggéré l'incorporation des facteurs démographiques dans le stéréotype des modèles cognitifs.

- Les connaissances de l'utilisateur sur le système de recherche qu'il utilise, car s'il connaît bien ce système il saura mieux définir son besoin.

- Les connaissances de l'utilisateur sur les tâches qu'il va essayer d'accomplir lors de la recherche. Une tâche est un ensemble d'actions accomplies par l'utilisateur dans le but d'atteindre un objectif déterminé. La connaissance des tâches à effectuer peut avoir un effet important sur la performance de la recherche.

- Les connaissances de l'utilisateur sur le domaine et le sujet qu'il cherche. La connaissance du domaine dépend du niveau d'expertise de l'utilisateur dans le domaine, car un débutant ne saura généralement pas formuler son besoin informationnel en utilisant les termes adéquats aussi bien

qu'un expert. Ceci peut influencer sur la pertinence des réponses fournies par le système.

En étudiant les connaissances des utilisateurs de DOC'INSA selon cette approche, il apparaît évident que nous pourrions appréhender les différentes caractéristiques des stéréotypes, en capter les propriétés et les comportements. C'est pourquoi, nous retenons cette approche fondée sur la description des connaissances de l'utilisateur.

De plus, en explorant les processus cognitifs, ALLEN a conclu que la conception des systèmes de recherche d'information doit tenir compte des transitions effectuées par l'utilisateur d'une étape de recherche à une autre.

2.2.3. Apports de l'IA

On constate que les actions accomplies par l'utilisateur, mises en évidence par l'approche sociologique, sont issues de ses connaissances, définies par l'approche cognitive. Approfondir et formaliser les modèles qui en découlent, pour exploiter et concevoir un système de recherche répondant aux vraies attentes de l'utilisateur et intégrant ses différentes caractéristiques, est un des buts des démarches des spécialistes de l'IA. Par exemple, PITRAT [PIT 90], envisage un modèle intégrant les connaissances sur les individus, qui peuvent être résumées ainsi :

- Savoir ce que l'individu sait ;
- Savoir ce qu'un individu peut faire ;
- Savoir comment un individu effectue une tâche ;
- Connaître les habitudes d'un individu.

Cependant, les connaissances de l'utilisateur ne sont pas statiques. Elles évoluent grâce à l'acquisition de l'expérience. Les tâches accomplies lors d'une recherche documentaire suivent aussi cette évolution. C'est pourquoi les modèles utilisateurs doivent intégrer de l'apprentissage afin de gérer l'évolution de l'utilisateur. C'est ce que tentent de faire les systèmes interactifs dits "intelligents", qui intègrent l'apprentissage automatique, que nous allons voir plus en détail par la suite (§2.4.4, Classification et mesure de similarité). Nous retenons cette approche utile pour maintenir et faire évoluer notre système, quelque soit le type de l'utilisateur.

Donc pour résumer, notre approche sera fondée sur la représentation et la formalisation des connaissances de l'utilisateur. Ces connaissances seront utilisées pour la construction des stéréotypes et exploitées en permanence pour gérer l'évolution du profil utilisateur, et ainsi garantir la pérennité de notre système.

2.3. Modèles formels de représentation d'un profil

On distingue deux grands groupes de représentation de profils [MID03]:

- Profil fondé sur la notion d'indice
- Profil fondé sur la représentation par le contenu.

La première représentation mémorise les indices des items définis comme valables pour chaque utilisateur. Ensuite des techniques de corrélation peuvent être utilisées pour trouver des utilisateurs similaires. Un exemple d'une telle représentation est le feedback d'intérêts. L'utilisateur a, à sa disposition, une échelle d'indices pour indiquer si le choix proposé par le système le satisfait ou non. La représentation du retour de pertinence est alors un ensemble d'items recommandés, avec des valeurs d'intérêt associées pour chaque utilisateur. Le problème de cette représentation est qu'elle demande trop d'effort en temps et en investissement à l'utilisateur qui la plupart du temps n'attribuera pas de valeurs.

Le deuxième groupe concerne la représentation par le contenu d'items spécifiques désignant l'intérêt d'un utilisateur particulier. Dans ce cas, les techniques d'apprentissages des machines peuvent trouver les items similaires. L'analyse fondée sur le contenu est souvent effectuée sur des documents textuels. Dans ce groupe on trouve plusieurs types de représentations : la représentation vectorielle de fréquence de termes, la représentation par classe binaire, ou par attribut valeur, ou par un profil multi-classe utilisant une ontologie, ou la représentation fondée sur la connaissance. La représentation vectorielle de fréquence de termes indique le nombre de termes apparaissant dans un document.

La représentation par classe binaire considère les intérêts de l'utilisateur comme un ensemble d'exemples positifs et négatifs. Chacun de ces groupes d'exemples est représenté par une collection de vecteurs de fréquence de termes que l'utilisateur a évalué. L'alternative de la classe binaire est l'ontologie des classes qui peut être créée comme une cartographie du domaine des concepts de l'utilisateur.

La représentation par la connaissance consiste en l'insertion de faits concernant l'utilisateur dans une base de connaissances à partir de laquelle des inférences peuvent être déduites afin de connaître les stéréotypes et les intérêts de l'utilisateur [MID03]. Notre choix de représentation dépendra du choix de la méthode de définition de profil qui va supporter cette représentation.

2.4. Méthodes de définition et d'évolution d'un profil

Il s'agit ici de s'intéresser à la définition et à l'évolution d'un profil pour un utilisateur (ou une communauté d'utilisateurs) donné, dans un environnement technologique précis. A partir des ou du formalisme retenu, plusieurs approches de définition de profils peuvent être retenues. Cependant nous allons exposer tout d'abord les techniques de raisonnement utilisées dans l'IA par les systèmes experts [BEN93] [GIU92]. L'idée est de pouvoir faire converger les différentes typologies dans un raisonnement artificiel, utilisable par les machines pour la résolution du problème.

2.4.1. Raisonnement par les règles

Depuis déjà 30 ans, les systèmes experts ont développé des outils fondé sur les connaissances et raisonnant par règles « si...alors... »¹. Ils ont un certain nombre de limites comme la forte composante procédurale, ainsi que l'absence de la séparation entre les connaissances déclaratives et procédurales.

2.4.2. Raisonnement fondé sur modèles

Ensuite, est apparu le raisonnement fondé sur les modèles et qui s'appuie sur une formulation sous-jacente des connaissances **sous forme de contraintes** (<http://sic.epfl.ch/SA/publications/FI94/10-94-page3.html>). Un modèle est dans ce cas représenté sous la forme d'un ensemble d'équations transformées en algorithmes et il peut répondre à des tâches complexes comme le diagnostic, le monitoring ou la configuration. L'identification des solutions admises par un tel système s'effectue grâce à des *moteurs de satisfaction de contraintes*.

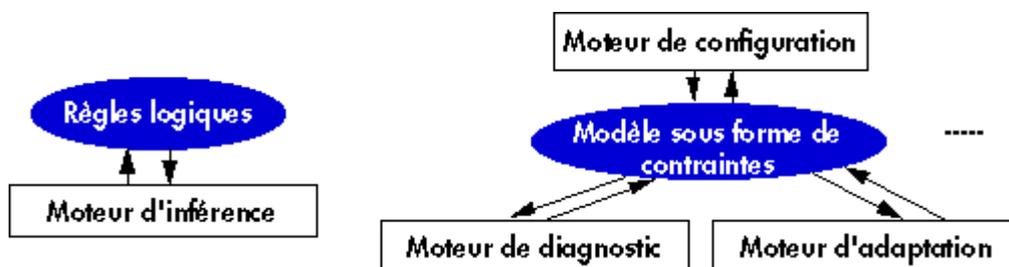


Figure 2 Raisonnements fondés sur les règles et les modèles

Cette méthode a l'avantage de permettre au système de se mettre à jour plus facilement que dans le cas précédent, et aux procédures spécialisées de se mettre en place selon les stratégies utiles pour les tâches de raisonnement

¹ <http://sic.epfl.ch/SA/publications/FI94/10-94-page3.html>

La limite est la complexité exponentielle qui peut être générée durant la résolution des problèmes de satisfaction des contraintes.

Pour palier cette complexité, une des solutions est de traiter le problème à un **niveau d'abstraction** plus élevé. Le choix du niveau d'abstraction réduit la complexité à un niveau voulu permettant de choisir des abstractions plus au moins fines en fonction des endroits où cela rapporte le plus. L'inconvénient de cette technique est qu'il faut s'accommoder de l'absence d'accès à toutes les solutions du problème original.

Mais ces techniques s'appliquent là où les modèles ou leurs fragments sont connus. Par contre, dans notre cas d'étude nous ne connaissons pas de modèles préétablis. C'est pourquoi, il serait nécessaire de chercher à procéder par des heuristiques résumant le savoir d'un expert et qui généralisent les expériences. Mais nous ne possédons pas non plus ce savoir. De plus, avec ces méthodes traditionnelles, il reste toujours très difficile d'acquérir la connaissance générale. C'est pourquoi notre choix porte sur une autre théorie, celle fondée sur la mémoire et sur le raisonnement humain, selon laquelle toute conclusion provient directement des analogies entre les expériences précédentes et le problème courant. Ainsi, la généralisation se fait précisément quand le problème concret doit être résolu. Il s'agit donc du **raisonnement à partir des cas**. Grâce à cette construction, les cas sont capables de tenir compte d'un contexte beaucoup plus large que les règles heuristiques. Nous allons détailler dans la suite cette approche par les cas.

2.4.3. Raisonnement fondé sur les ontologies

Comme dans d'autres domaines de l'ingénierie des connaissances, des travaux ont porté sur l'intérêt de l'utilisation des **ontologies** pour la modélisation de l'utilisateur. Par exemple, Judy Kay [KAY99] a travaillé sur le sujet de l'utilisation des ontologies pour créer des modèles utilisateurs réutilisables et scrutables en préconisant que le modèle utilisateur a besoin d'une ontologie agréée et d'une représentation, afin d'être ainsi utilisable par différents programmes de l'application. Elle met l'accent sur le fait qu'un puissant système de personnalisation nécessite un modèle réutilisable. La réutilisabilité a de l'importance car la maintenance des modèles de l'utilisateur induit un coût élevé en temps et en moyens. Ces idées ont été mises en pratique dans le système « um » cités dans la partie concernant les outils de référence de modélisation de l'utilisateur [KOB01]. Dans le cadre de cette technique, le domaine est modélisé comme une structure taxonomique de concepts en relations entre eux et chacun ayant ses attributs. C'est de cette manière que nous allons construire notre modèle de connaissances du domaine de recherche d'informations disponibles dans la bibliothèque numérique des thèses. Mais Kay, va plus loin dans son raisonnement. La fonctionnalité de scrutabilité données par l'ontologie doit permettre aux utilisateurs d'analyser minutieusement leurs modèles. Cela peut être particulièrement utile quand un utilisateur exprime un certain nombre de préférences comprises dans un modèle plus large.

2.4.4. Raisonnement à Partir des Cas

Les fondements

« Le raisonnement à base de cas est une approche de résolution de problèmes qui utilise des expériences passées pour résoudre de nouveaux problèmes. L'ensemble des expériences forme une base de cas » [LAM02] Les Fondements du CBR (Case-based Reasoning) traduction anglaise de RaPC, proviennent de travaux en sciences cognitives menés par Schank et son équipe de recherche pendant les années 80. Ils ont développé la théorie de la mémoire dynamique, selon laquelle les processus cognitifs de compréhension, de mémorisation et d'apprentissage utilisent une même structure de mémoire représentée à l'aide de schémas de représentation de connaissance tels que des graphes conceptuels et des scripts.

Pour citer un certain nombre d'auteurs, le CBR est un raisonnement par mémorisation

(LEAKE, 1996). Un raisonneur fondé sur les cas, résout de nouveaux problèmes en adaptant les solutions déjà utilisées pour résoudre d'anciens problèmes [SCH89]. [KOL93] explique que le raisonnement à partir des cas est à la fois la manière dont les humains utilisent les expériences pour résoudre des problèmes et la manière dont on peut obliger les machines d'utiliser les cas. Plus précisément, un système fondé sur le CBR est défini par une combinaison de *processus* et de *connaissances* sur le *cas* que nous allons étudier plus largement par la suite.

Les connaissances

Elles portent sur l'indexation, la base de cas, les mesures de similarité et l'adaptation [LAM02].

Le vocabulaire *d'indexation* est un ensemble d'attributs ou de traits caractérisant la description des problèmes et de solutions du domaine. La *base de cas* est un ensemble d'expériences structurées, à exploiter durant les phases de recherche, d'adaptation et de maintenance. Les *mesures de similarité* sont des fonctions pour évaluer la similarité entre deux ou plusieurs cas. Les *connaissances d'adaptation* sont des heuristiques du domaine permettant de modifier les solutions et d'évaluer leur applicabilité à de nouvelles situations.

Description du cas

Sur le site de la communauté internationale qui effectue des recherches très actives au sujet du RaPC (<http://www.ai-cbr.org/>), il est expliqué qu'un cas peut décrire soit des tâches analytiques afin de classer, diagnostiquer ou pré fabriquer des suggestions de solution, soit des tâches synthétiques pour générer des plans complexes afin d'arriver à la solution (<http://demolab.iese.fhg.de:8080/help-tasktype.html>¹). Les besoins de l'étude, à savoir l'analyse du modèle de l'utilisateur afin de définir une requête de recherche satisfaisant ses besoins, nous placent plutôt dans les tâches analytiques.

Elles répondent à 3 types de problèmes à résoudre : ceux de la classification, du diagnostic, du support de décision. Les tâches de classification visent à classer les objets ou les situations dans des ensembles donnés de classes. Les tâches de diagnostic interviennent quand il manque de l'information pour pouvoir procéder à la classification. Donc, le diagnostic inclut, en plus des tâches de classification, la sélection du meilleur test qui acquière l'information manquante d'une manière avantageuse, rapide et sûre. Les tâches concernant le support de décision s'adaptent quand il y a une forte interaction entre la machine et l'utilisateur. Le but de la classification ou du diagnostic est défini ou redéfini au dernier moment, durant la résolution du problème.

Typiquement, un cas *décrit* une situation diagnostique et contient en général la description des symptômes, du défaut et sa cause, ainsi que la description de la stratégie de réparation.

Modélisation du cas

Aujourd'hui on trouve trois types de modélisation du cas : structurel, conversationnel et textuel [LAM02].

Le choix de ces modèles dépend des conditions du domaine et de la structure des données disponibles dans le cas.

Dans les cas de *structure* simple, la représentation se fait par une liste d'*attributs-valeurs*, d'autant plus qu'elle est facile à stocker et à récupérer dans le système de CBR. La représentation par un *modèle* orienté objet est utilisée dans les cas de manipulation d'objets structurés et complexes. Le cas est une *collection d'objets*, c'est-à-dire des instances de classes dans le sens de la programmation orientée objet. Dans des situations plus spécifiques, on peut utiliser soit les *graphes* où les cas sont des ensembles de noeuds et d'arcs, soit les *vecteurs* où le cas est un seul ou un ensemble de *vecteurs* se trouvant dans l'espace de la base des cas, soit les plans où les cas sont des

¹ Lien du site <http://www.ai-cbr.org/tools.html>

ensembles ordonnés d'actions, ou encore des prédicats logiques où le cas est un ensemble de formules atomiques.

Les applications commerciales du CBR ont connu plus largement le *modèle conversationnel* qui met en premier plan l'interaction entre l'utilisateur et le système. Il définit progressivement le problème à résoudre et concerne trois parties : la description textuelle du problème, des questions et des réponses, servant d'index et permettant d'obtenir plus d'informations sur la description du problème, et une action décrite textuellement fournissant la solution à mettre en oeuvre pour le problème.

Le modèle *textuel* comme son nom l'indique, les expériences sont décrites et contenues dans des documents textuels non structurés ou semi-structurés.

Nous allons utiliser la représentation du cas par une liste d'attributs – valeurs, car elle apparaît suffisante pour les besoins de notre système et facile à mettre en oeuvre.

Le processus

Le processus (**Figure 3**) est constitué de quatre étapes : la construction, la recherche, l'adaptation et la maintenance [LAM02].

La *construction* est une première étape, souvent réalisée manuellement par le concepteur du système nécessitant une très bonne connaissance du domaine pour stocker des cas de départ significatifs, à partir du cadre applicatif pour la sélection du vocabulaire d'indexation et la définition des métriques de similarité.

La *recherche* détermine les cas de la base de cas les plus similaires au cas à résoudre. Elle est habituellement implémentée par la méthode des plus proches voisins ou par la construction d'une structure de partitionnement par induction. Dans le premier cas, des métriques de similarité sont utilisées pour mesurer la correspondance entre chaque cas de la base et le nouveau problème à résoudre. Dans le deuxième cas, un arbre est généré pour répartir les cas selon différents attributs et ainsi guider le processus de recherche.

L'*adaptation* modifie les réponses des cas retrouvés pour construire une nouvelle solution. Deux approches de mise en oeuvre sont appliquées, la méthode transformationnelle et la méthode générative. Dans la première, on modifie des solutions antécédentes en les réorientant pour satisfaire le nouveau problème, et dans la deuxième, on garde les traces des étapes qui permettent de générer la solution de chaque cas traités et on applique une de ces étapes dans une nouvelle solution, pour un nouveau cas. Dernièrement des travaux [LAM02] ont tenté de combiner les 2 solutions. L'adaptation peut être complètement automatique, ce qui est assez rare, ou semi-automatique grâce à l'intervention humaine.

La *maintenance* assure l'intégration des nouvelles solutions dans la base des cas et la modification des structures du système pour en optimiser les performances. Plusieurs approches peuvent être mises en oeuvre : insérer un cas dans la base, modifier la structure de la base pour faciliter l'exploitation ou encore modifier les attributs des cas et leur importance relative (ou poids). Cette intégration de nouvelles solutions dans la base correspond à l'acquisition de nouvelles expériences pour l'être humain et donc, à l'activité d'apprentissage. C'est pourquoi, les techniques d'apprentissage par les machines doivent être adaptées dans ce processus.

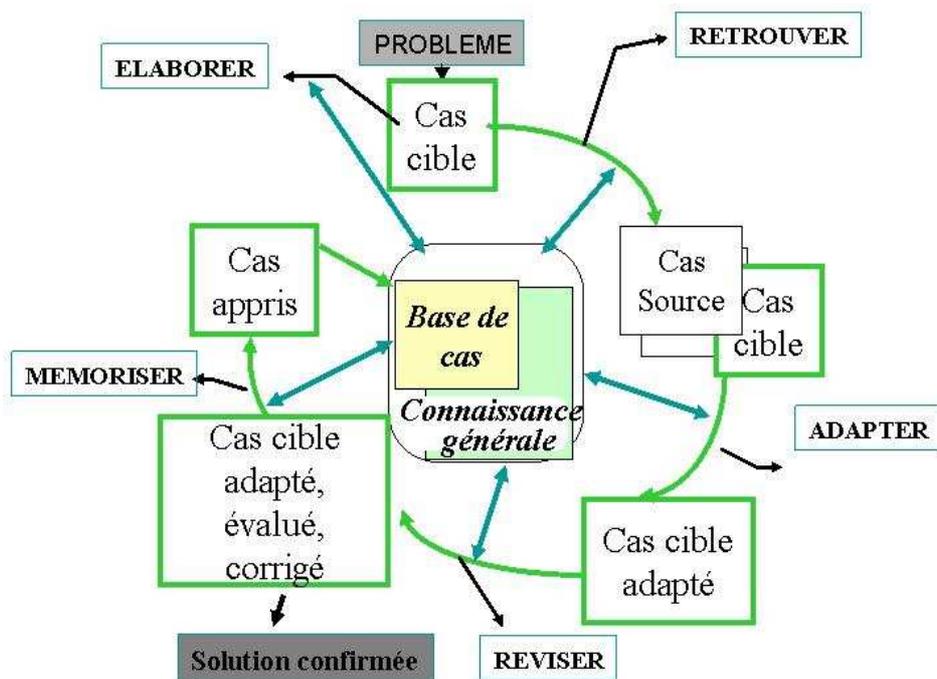


Figure 3 . Processus de CBR¹

2.4.5 Classification et mesure de similarité

Le point crucial du raisonnement à partir du cas, est l'efficacité du choix du cas similaire qui va permettre l'exploitation d'une ancienne expérience pour le traitement d'un nouveau cas. Concernant les tâches analytiques décrites par les cas, nous avons souligné qu'un premier problème à résoudre est la classification des cas. Plusieurs algorithmes de classifications existent dans le cadre des techniques de l'apprentissage par les machines (machine - learning techniques) [MID03].

Ces techniques sont divisées en deux groupes : celles d'apprentissage non supervisé, et celles d'apprentissage supervisé. Les techniques supervisées prennent un ensemble de cas catalogués comme exemples pour la catégorisation des nouveaux cas. Cet ensemble s'appelle également ensemble d'entraînement. Les techniques non supervisées n'utilisent pas d'ensembles d'entraînement, car les classes sont générées par des modèles (patterns) à partir d'exemples non catégorisés.

Certes, catégoriser des ensembles d'entraînement est une tâche qui prends du temps et s'effectue le plus souvent à la main, mais l'avantage des techniques supervisées est la fiabilité, puisqu'elles évitent les erreurs introduites par la génération automatique des classes. De plus, dans notre étude, nous ne disposons pas de modèles pour la catégorisation des cas. Pour ces raisons, nous nous intéressons aux techniques supervisées de l'apprentissage par les machines.

Parmi les techniques supervisées, on peut envisager les méthodes fondées sur l'instance. La méthode la plus communément utilisée dans la littérature des systèmes de recommandation pour la modélisation de l'utilisateur est l'algorithme du k-plus proche voisin. Un autre point positif de cet algorithme est qu'il a été largement utilisé dans les implémentations du raisonnement à partir du cas, ainsi que la résolution pour d'autres problèmes de classification. De plus il paraît tout à fait adapté à notre domaine. Nous avons donc porté notre choix sur cet algorithme.

Plus précisément, la règle de classification du k-plus proche voisin (k-Nearest Neighbors) est un modèle de classification statistique non paramétrique. Il est statistique, car il utilise des mesures statistiques dans ses calculs, et non paramétrique, car il ne prend pas en compte une fonction de

¹ http://liris.cnrs.fr/%7Eamille/enseignements/master_ia/Alain/rapc_session3_cycle_raisonnement.html

distribution probabiliste d'un échantillon donné [FUJ96] [PEK04][FRI94].

Après avoir défini le champ d'action de ce modèle, nous allons voir comment il fonctionne. L'idée est de trouver des cas d'apprentissage similaires au nouveau cas, en faisant "voter" un nombre fixe de voisins de ce cas. Pour cela, on doit choisir la classe de solution prédite pour ce cas et aussi le cas source de cette solution qui devrait être plus facile à adapter. Donc, il faut établir une règle de décision du classificateur, et aussi une valeur seuil qui doit séparer les classes. Ainsi on aura un certain nombre de classes de cas appelés ensembles d'entraînement. La règle de décision est établie par le calcul de similarité entre les cas, qui revient également à calculer la distance entre les cas. La mesure de similarité le plus souvent utilisée est fondée sur la distance euclidienne entre deux cas et se formalise comme suit :

$$d = \frac{\sum p_i \times d_i}{\sum p_i}$$

Où p_i est le poids d'un descripteur du cas, d_i la distance entre deux mêmes descripteurs de différents cas et i indique le rang du descripteur au sein du cas qui va de 1 au nombre maximum de descripteurs présents dans le cas.

A la suite, nous présentons l'algorithme du calcul du cas plus proche voisin.

Algorithme de calcul du cas plus proche voisin.

Début

{

Initialiser la valeur de T // T est le seuil de similarité minimum pour être considéré comme voisin de ce cas cible en toute théorie, il pourrait y avoir autant de seuils différents que de cas cibles

Initialiser la valeur de K // K représente le nombre de cas voisins nécessaire et suffisant pour qu'une classe de solution soit éligible comme candidate pour la solution cible

Initialiser liste_cas_voisins // la liste est vide au départ

Tant_que (il reste des cas à comparer)

{Comparer la description du cas cible avec la description du cas source suivant de la base de cas;

Si (la similarité cas_cible/cas_source > T)

Alors (Insérer le cas source dans liste_cas_voisins);

Passer au cas suivant;

} Fin_Tant_que

Choisir la classe solution éligible la plus représentée dans liste_cas_voisins;

Si (aucune classe de solution éligible)

Alors (conclure au manque d'expérience suffisante pour ce type de cas cible);

FIN;

Choisir le cas source le plus similaire de la classe solution élue;

}

Fin

Après avoir étudié les concepts principaux du raisonnement à partir des cas, nous allons résumer les avantages et les désavantages de cette technique

2.4.6. Avantages et limites du RaPC

Avantages

Tout d'abord, l'*effort de l'acquisition* de la connaissance générale est réduit, car le système est fondé sur la connaissance du cas qui est facile à percevoir puisque il correspond à une expérience réelle et bien précise. Ensuite, la **démarche de traitement** des cas et de résolution du problème se rapprochant de la démarche naturelle d'apprentissage des humains, le système peut lui-même développer la performance de résolution du problème à travers la réutilisation, en utilisant seulement les données existantes, en s'améliorant ainsi dans le temps et en s'adaptant aux changements de l'environnement. Enfin, la *maintenance* d'un tel système demande moins d'effort car les cas sont indépendants les uns des autres et donc on peut en ajouter ou supprimer facilement sans pour autant perturber tout le système. Ces cas peuvent également être facilement compris par tous les utilisateurs, que ce soient des débutants ou des experts.

Désavantages

Une première limite de cette approche est la difficulté de démarrer le mécanisme de RaPC, car au départ il n'y a pas un **nombre suffisant de cas** pour pouvoir réaliser une recherche par similarité et le système n'aura pas à sa disposition une base de cas assez conséquente pour les calculs. Une solution consiste à s'appuyer sur les stéréotypes de profils et leurs caractéristiques de recherche, à la phase de démarrage.

La limite principale porte sur le problème de **l'efficacité du cas retrouvé**. S'il est inapproprié, le système n'est pas aidé pour trouver la solution du problème. Cela peut arriver si le mécanisme de recherche faillit dans sa tâche de récupération du meilleur cas, ou plutôt de cas le plus similaire, ou s'il n'y a pas de bons cas disponibles dans la bibliothèque des cas. La solution serait de proposer des cas par d'autres moyens que par le CBR afin d'enrichir suffisamment la base pour pouvoir mettre en route les calculs de similarité. C'est pourquoi un axe de notre travail serait de trouver un modèle approprié de cas facilement adaptables à une base de cas peu riches, afin de mettre en route le plus vite possible les mécanismes de CBR. Mais dans un premier temps nous allons nous appuyer sur l'utilisation des stéréotypes, puis nous allons nous attacher à adapter la technique adéquate de calcul de similarité pour notre système.

2.5. Historique des systèmes de modélisation de l'utilisateur

Les premiers travaux sur la modélisation de l'utilisateur datent de la fin des années '70 avec les propositions de ALLEN en 1979, de Cohen et Perrault en 1978 ainsi que d'Eliane Rich en 1979. Pendant une dizaine d'années, dans les systèmes applicatifs représentant les modèles utilisateurs, il n'y avait pas de distinction claire entre les composantes servant les finalités du modèle utilisateur, et les composantes destinées aux autres tâches. A partir du milieu des années '80, les travaux de KOBASA, ALLGAYER et d'autres mettent en évidence cette distinction, mais ils ne mettent pas en place la réutilisation des composantes du modèle de l'utilisateur pour d'autres systèmes. Arrive alors la fin des années '80 où voit le jour le GUMS (General User Modeling System) par Tim Finin [KOB01]. Il met en place une hiérarchie simple de stéréotypes, et des facettes Prolog décrivant les membres de chaque stéréotype et les règles de raisonnement du système vis à vis de ces membres. Il définit également un moteur d'investigation de cohérence entre les nouvelles facettes et celles existantes. Il a servi comme un framework pour de simples fonctionnalités des systèmes qui lui ont succédé. Ensuite KOBASA introduit la notion des « Systèmes Shell » pour ce type d'applications, s'appuyant sur les éléments suivants :

- La représentation des **assomptions selon les caractéristiques** de chaque **utilisateur** comme les connaissances, les erreurs, les buts, les objectifs, les préférences, les tâches et leurs capacités.
- La représentation de caractéristiques communes d'utilisateurs comme les **stéréotypes**, les groupes ou sous-groupes etc.

- La mémorisation du comportement des utilisateurs comme les interactions passées enregistrées par le système, ou encore la création d'assomptions fondées sur l'interaction.
- La généralisation d'interactions stockées dans les historiques de plusieurs utilisateurs pour créer des stéréotypes.
- Les moteurs d'inférence qui permettraient de dessiner de nouvelles assomptions fondées sur des assomptions initiales, de fournir l'assomption en cours de l'utilisateur selon une justification, ainsi que l'évaluation des nouvelles entrées dans le modèle de l'utilisateur courant et sa comparaison avec des standards donnés.
- La maintenance continue du modèle de l'utilisateur.

Ensuite, une autre génération a vu le jour : les applications client-serveurs fournissant des avantages comparables avec les composantes embarquées du modèle de l'utilisateur. Cependant, un certain nombre de caractéristiques communes se dégagent de ces outils :

- L'information sur l'utilisateur est maintenue dans un répertoire central ou distribué virtuellement et mis à disposition de plusieurs applications en même temps.
- Les informations de l'utilisateur acquises par une application sont utilisées par d'autres.
- L'information sur l'utilisateur est stockée dans un emplacement non redondant. Ainsi, la consistance et la cohérence de l'information donnée par différentes applications, peut être atteinte plus facilement.
- L'information sur les groupes d'utilisateurs, ou bien disponible sous formes de stéréotypes tels que développés par RICH [RICH1979], [RICH1983], ou même calculée dynamiquement comme des modèles de groupes d'utilisateurs, peut être maintenue avec une basse redondance.
- Les méthodes et les outils du système de sécurité, d'identification, d'authentification, du contrôle d'accès et de l'encryptage, peut s'appliquer pour la protection des modèles des utilisateurs dans les serveurs d'applications.
- Et également, l'information sur l'utilisateur, dispersée dans l'entreprise peut s'intégrer plus facilement dans le réceptacle du modèle de l'utilisateur.

Pour finir, on peut remarquer que la gestion de l'information sur l'utilisateur selon des groupes, ou modèles de groupes d'utilisateurs, ou encore selon les stéréotypes, a pris une place importante dans le développement des applications pour la modélisation de l'utilisateur.

Chapitre 3 : La modélisation de l'utilisateur – Notre proposition

Dans notre contexte qui concerne les bibliothèques numériques de thèses scientifiques, nous explorons diverses pistes pour proposer un accès pertinent au contenu de ces documents.

Un premier axe de recherche nous a conduit à définir un modèle de thèse fondé sur l'intégration de balises sémantiques dans le document [BOH05][ABA05] à partir duquel on pourra extraire les parties de thèses correspondant le mieux à un concept ou à une thématique recherchée[ABA05].

Grâce à une collaboration avec la bibliothèque DOC'INSA de l'INSA, nous avons travaillé sur un corpus de documents disponibles pour les aspects expérimentaux. En effet, DOC'INSA a mis en place depuis 1997 le projet CITHER (Consultation en texte Intégral des Thèses en Réseau), qui permet la diffusion par Internet et l'accès aux thèses scientifiques. Actuellement, l'utilisateur de cette bibliothèque numérique peut accéder au contenu d'une seule thèse à la fois, sans pouvoir obtenir des extraits pertinents correspondant à une unité de corpus plus fine que le chapitre. Les raisons de ce mode de consultation sont l'utilisation d'un format inadéquat à la recherche de l'information, tel que PDF, la description du contenu uniquement par les mots-clé ajoutés extérieurement aux documents, et l'utilisation seulement des métadonnées Dublin Core qui apportent des informations générales sur la thèse.

Nous avons proposé de mettre en place un système remédiant à ce problème avant tout, grâce aux métadonnées insérées en tant que « tags sémantiques » dans le corpus de chaque thèse. C'est l'auteur de la thèse qui va pouvoir insérer des tags identifiant les concepts qu'il définit dans son document qui à son tour obéit à un modèle de document défini grâce à la technologie XML Schémas.

Le second axe, qui fait l'objet de ce sujet de master, porte sur la définition et l'exploitation du profil utilisateur en vue de satisfaire au mieux la demande de l'utilisateur dans ce contexte de Recherche d'Information. Le but est de corriger les défauts du système actuel en laissant la possibilité d'accéder à plusieurs thèses à la fois en obtenant des extraits pertinents correspondant à une unité de corpus plus fine que le chapitre. On doit également prendre en compte l'aspect personnalisation du système de recherche d'information pour l'accès pertinent au contenu des thèses scientifiques de CITHER. Le but est de pouvoir donner des réponses pertinentes et adaptées à l'utilisateur même quand il est dans l'incapacité de bien préciser sa requête, autrement dit, de faciliter l'écriture des requêtes de l'utilisateur en les enrichissant avec des données invariantes, communes à toutes les requêtes, mais spécifiques à un utilisateur, ou un groupe d'utilisateurs.

C'est pour cette raison que dans ce chapitre nous allons décrire précisément la démarche suivie pour définir la typologie des connaissances retenues qui représentent notre profil de l'utilisateur, sa formalisation et sa représentation ainsi que la manière de faire évoluer ce profil.

A partir de l'étude des typologies des connaissances des utilisateurs, nous avons construit un modèle de connaissances qui pourrait caractériser différents stéréotypes. Ensuite ce modèle nous a servi pour décrire, construire et indexer le cas, qui d'ailleurs conjugue les caractéristiques des stéréotypes. Le cas va traduire formellement une expérience de l'utilisateur. En accumulant les cas, nous allons pouvoir suivre l'évolution du profil d'un utilisateur, mais également les tendances de comportement d'un groupe d'utilisateurs, ou des différents stéréotypes présents dans le système.

3.1. Typologie de connaissances impliquées

Au cours d'une session de recherche, l'utilisateur exprime des intentions ou des besoins de recherches spécifiques, liés à ses intérêts et ses connaissances qui sont représentés par des mots clés ou des documents intéressants pour l'utilisateur, ou encore, ils sont déduits à partir de son comportement. L'expérience montre qu'il est difficile d'anticiper toutes ces caractéristiques afin de

l'assister et de lui apporter l'aide nécessaire dans tous les cas et les contextes possibles (§2.5. Historique des systèmes de modélisation de l'utilisateur). Cependant, pour notre étude nous avons dégagé un certain nombre de connaissances qui nous paraissent les plus importantes et qui sont organisées en cinq groupes : connaissances générales, connaissances du domaine, connaissances du système, connaissances de recherche et connaissances de restitution (Figure 4)

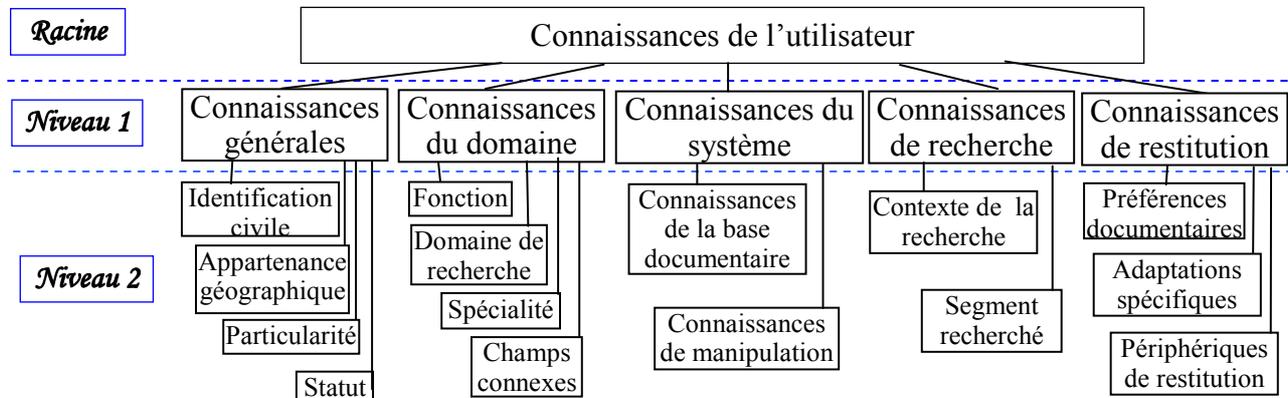


Figure 4. Modèle de l'utilisateur fondé sur les connaissances

3.1.1. Connaissances générales CG (Figure 4)

Elles sont liées à l'identification civile de l'utilisateur (civilité, nom, prénom), à l'appartenance géographique (adresse, ville, pays), éventuellement à des particularités de l'utilisateur comme le handicap (visuel, surdité...), et à l'appartenance socio-culturelle qui ici sera traduite par le statut de l'utilisateur. Il s'agit du statut qui peut prendre des valeurs comme "constructeur", "bibliothécaire", "administrateur informatique" "Invité". Le constructeur est un utilisateur susceptible de participer à la construction (saisie, direction, correction) d'une thèse. Un utilisateur extérieur, qui n'a jamais eu d'identification à l'INSA, mais qui se connecte seulement sur le module de recherche d'information a le statut "d'invité". Cet utilisateur ne sera jamais reconnu autrement que par ce statut. Toutes des données sont supposées rester immuables.

3.1.2. Connaissances du domaine CD (Figure 4)

Ce groupe contient les connaissances liées à la fonction de l'utilisateur (enseignant, étudiant, enquêteur), le domaine de recherches scientifiques et la spécialité de l'utilisateur. Nous intégrons également des domaines connexes d'applications. Par exemple, un doctorant qui fait des études en mathématiques et qui travaille sur les calculs automatisés, va avoir une fonction d' « étudiant » qui travaille sur les domaines des «mathématiques», spécialisation «calculs automatisés» et va avoir comme domaine d'application l'«informatique». Nous allons laisser la possibilité aux utilisateurs de choisir jusqu'à 3 domaines d'applications (nombre qui pourra évoluer selon les résultats).

Il est à remarquer dans ces deux typologies de connaissances que nous avons introduites, caractérisent différents types d'utilisateurs, mais la frontière est discutable. C'est pourquoi, avant de passer à la suite de la typologie, nous allons justifier le choix de ces deux groupes (CG et CD).

3.1.3. Caractérisation des stéréotypes

Le point litigieux réside dans l'établissement de la différence entre le "statut" rencontré dans les CG et la "fonction" des CD. Nous avons décidé de garder les deux appellations selon deux critères : (i) la frontière qui sépare les différents types d'appellations et la (ii) finesse des

caractéristiques qu'ils développent et qui peuvent nous aider à affiner les comportements des stéréotypes d'utilisateurs durant leurs sessions de recherche d'information.

Par exemple, de part l'exercice de leurs activités, on prend comme règle de gestion le fait qu'un constructeur, comme par exemple un professeur, ne sera sans doute jamais un bibliothécaire. Au contraire un étudiant thésard est susceptible de devenir un professeur correcteur. En pratique, durant une session de recherche d'information dans le système, le professeur va s'intéresser au contenu de la thèse, alors que la bibliothécaire va s'intéresser aux éléments d'archivage se trouvant dans les post - liminaires des thèses.

Ensuite, un constructeur et un invité (ce qui correspond au statut) restent figés dans leurs activités. A l'inverse, un enseignant peut être décliné en directeur, membre du jury, ou rapporteur, et un étudiant peut être un pré-doctorant, un doctorant ou un ingénieur. Ces groupes d'utilisateurs définis par la description de la fonction, peuvent changer d'activité. Grâce à ce point de vue, nous aurons gagné en finesse de définition des caractéristiques des stéréotypes intervenant dans notre système. Par exemple un doctorant pourrait être amené à chercher dans des parties de l'état de l'art des thèses, alors qu'un ingénieur pourrait s'intéresser plutôt aux prototypes développés pour les thèses. Ceci nous aidera à mieux définir et formaliser les comportements des utilisateurs appartenant à tel ou tel stéréotype.

C'est un travail de réflexion que nous allons continuer en rencontrant d'autres caractères de stéréotypes au fur et à mesure que nous développons les typologies des connaissances de l'utilisateur.

3.1.4. Connaissances du système CS (Figure 4)

Un utilisateur qui connaît bien le système, définit en principe mieux son besoin. Nous estimons qu'un utilisateur sera plus performant lors d'une session de RI s'il connaît la manière dont les thèses sont stockées dans la base documentaire, sans pour autant rentrer dans des détails techniques d'informaticiens, et en connaissant les fonctionnalités du système (Figure 5).

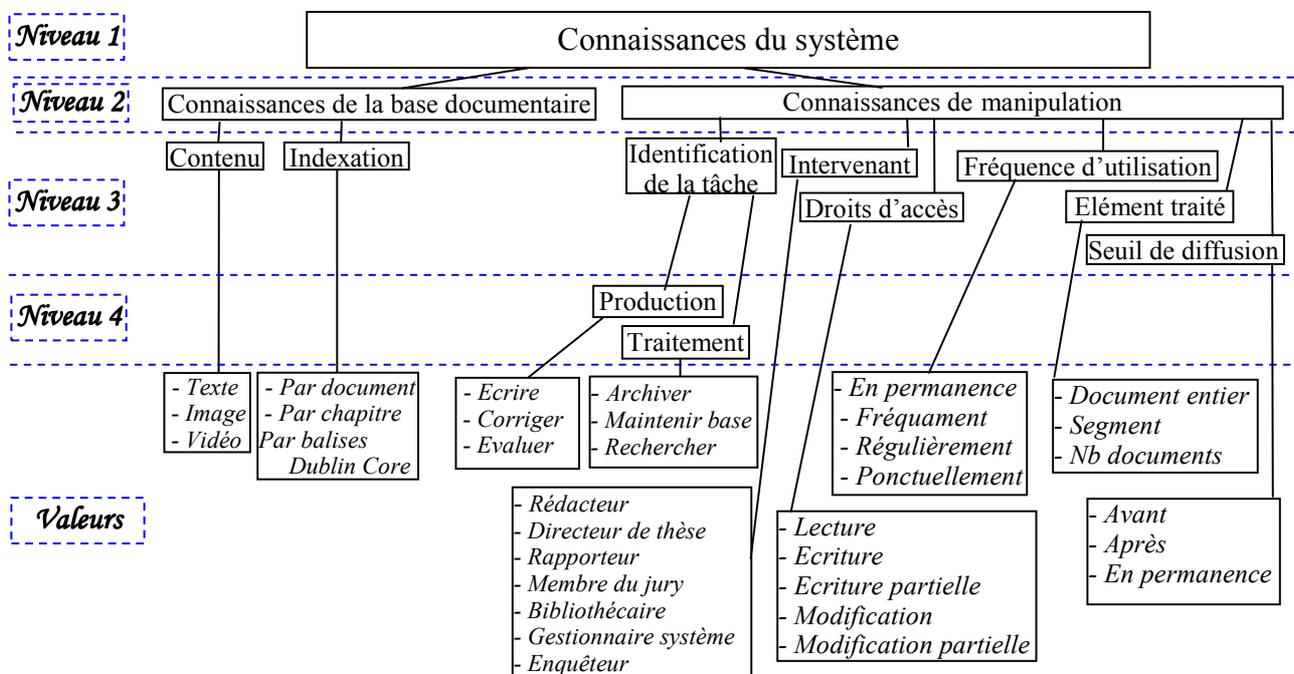


Figure 5. Modèle des connaissances de l'utilisateur sur le système

En définissant ces connaissances, l'idée est de déterminer les pré-requis que l'utilisateur a par rapport au système de RI, afin qu'il ait une idée avant la session de recherche de ce qui peut le limiter ou non. Par exemple, un simple utilisateur n'aura pas les mêmes connaissances du système qu'un utilisateur qui a déjà soutenue et déposée une thèse à l'INSA. Un docteur INSA sait que dans

ce système il ne peut pas consulter toutes les thèses soutenues durant l'année courante, mais seulement celles qui ont été diffusées, alors qu'un utilisateur occasionnel ne connaîtra pas ce détail.

Aussi, un docteur qui a travaillé avec l'outil actuel de DOC'INSA n'aura pas les mêmes connaissances du système qu'un docteur qui travaillera avec le nouveau système que nous développons. En effet la grande différence entre les deux réside dans le fait que dans l'ancien système les mots-clés d'indexation des thèses concernent les documents en général, ou au mieux les chapitres de quelques thèses, et non pas des extraits bien précis balisés par les auteurs qui sont experts du contenu de leur document. Ainsi, dans le système actuel, quand l'utilisateur effectue une recherche par mots-clés, il lui faudra lire tout un chapitre de thèse qu'il aura obtenu en résultat de recherche, et pour une seule thèse à la fois. Alors qu'avec la nouvelle version que nous avons définie pour CITHER il pourra accéder à plusieurs passages qui le concerneront exactement, et extraits de plusieurs thèses à la fois. Par contre, que ce soit dans l'ancien ou le nouveau système, les recherches seront effectuées seulement en format texte.

Donc, en général, en définissant cette typologie de connaissances nous avons gardé à l'esprit l'idée que l'utilisateur doit pouvoir répondre aux questions « qu'est-ce que je manipule? », « comment je manipule; » et « quand je manipule? ». Ainsi nous pourrons dégager encore d'autres caractéristiques et événements traités par les utilisateurs de différents stéréotypes. Dans tous les cas, tous les utilisateurs doivent connaître le type de contenu qu'ils manipulent, que ce contenu est indexé d'une certaine manière, qu'ils veulent effectuer une tâche bien précise, qu'ils ont un statut défini pour des aides possibles, qu'ils ont des droits bien spécifiés, utilisables quand ils le souhaitent dans n'importe quel élément d'une ou plusieurs thèses quand leur est permis.

Par exemple, un utilisateur cherchant de l'information manipule que du contenu texte qui peut être indexé par chapitre. Il effectue la tâche de recherche. S'il le souhaite, il consulte en permanence le système et il n'a que le droit de lecture seule sur une ou plusieurs parties des thèses déjà diffusées.

Une fois que l'utilisateur connaît les possibilités offertes par le système, il peut saisir sa requête de recherche que nous allons décrire par la suite.

3.1.5 Connaissances de recherche CRE (Figure 4)

Cela a déjà été précisé, l'utilisateur doit pouvoir procéder à des recherches plus au moins précises et le système doit pouvoir donner des réponses pertinentes, quelque soit le mode opératoire. C'est pourquoi nous avons choisi un certain nombre d'éléments qui pourront correspondre à ces différentes recherches et qui seront recueillis par le biais d'un écran de saisie.

Deux types de données vont se conjuguer pour définir la requête de l'utilisateur, le contexte de recherche et le segment recherché (**Figure 6**).

Le contexte de recherche concerne le type de document traité et la manière dont il va être traité. Par contre, le segment recherché indiquera l'élément à traiter dans le cadre d'une recherche précise. Les types de document correspondent aux documents disponibles à DOC'INSA (liste dans l'**Annexe 1**).

Le type de la recherche est un élément important du modèle des connaissances de recherche. C'est le plus imprécis et par conséquent le plus difficile à cerner.

Si l'utilisateur procède à une recherche par thème, le système va déclencher une recherche par domaines et sous-domaines. Afin de simplifier la formalisation des champs et de rester le plus fidèle possible à la réalité, nous avons décidé d'attribuer la liste des départements d'enseignements à l'INSA, à la liste des domaines, et la liste des spécialités à la liste des sous-domaines. Par exemple, le département de l'informatique dispense des cours de « Connaissances et raisonnement ». Donc, il s'agit du domaine informatique et du sous-domaine de connaissances et raisonnement (**Annexe 1**).

Si l'utilisateur procède à une recherche par sujet, le système va chercher sur la base des spécialités et des champs connexes. Pour les mêmes raisons que dans le cas de la recherche par thème, nous allons trouver la liste des spécialités parmi les groupes de recherche de l'INSA (ex.: groupes de travail de LIRIS). Les champs connexes proviendront des universités de co-tutelle s'ils ne correspondent pas au domaine principale, ou parmi les études de contenus réalisées sur un certain nombre de thèses numériques. Par exemple, la thèse intitulée « Consultation assistée par

ordinateur de la documentation en sciences humaines » a été réalisée au sein du groupe de travail « Données, Documents et connaissances » en co-tutelle avec une université d'archéologie en Grèce.

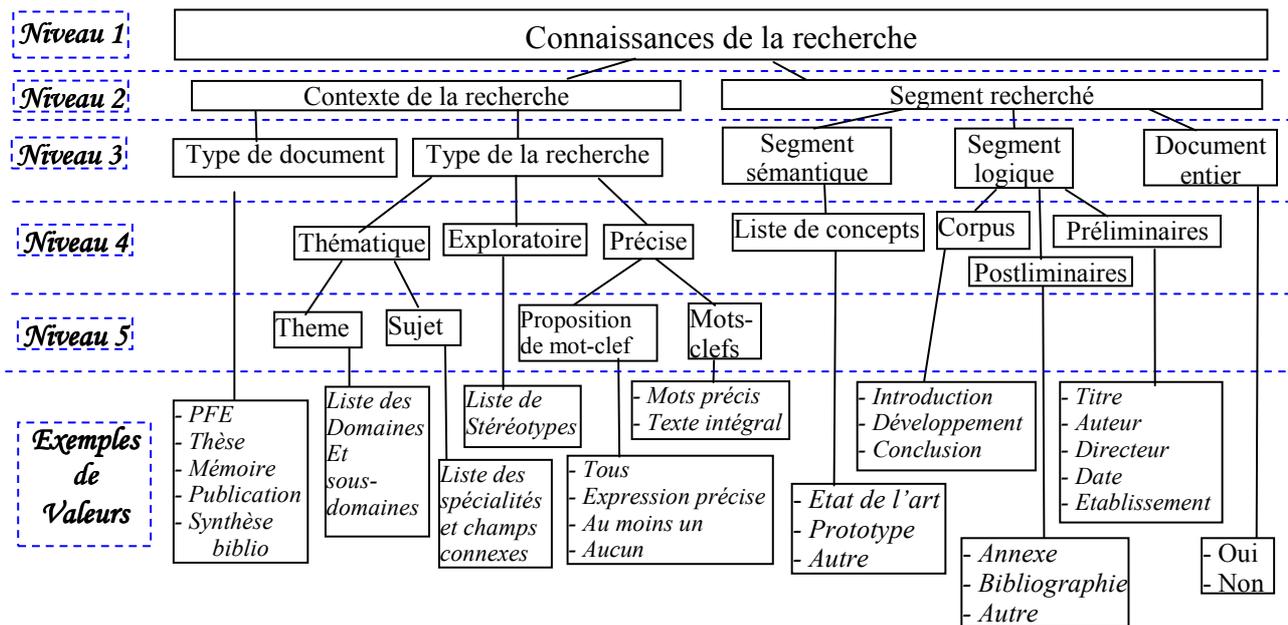


Figure 6. Modèle des connaissances de la recherche

Donc, sa spécialité est « Donnée, Document et connaissances » et son « champ connexe » l'archéologie.

Ce sont tous des éléments disponibles à l'utilisateur sur l'écran de saisie sous la forme de listes déroulantes où il pourra effectuer un choix, et par conséquent ceci deviendra une connaissance de l'utilisateur sur la recherche (**Annexe 1**).

Certes, cette manière de définir notre domaine est un peu restrictive, mais cela satisfait nos besoins actuels pour la première phase d'expérimentations. Dans le cadre des travaux futurs il est important de réaliser une étude pour la construction d'une manière scientifique de l'ontologie du domaine avec les sous-domaines et les spécialités.

Pour la recherche exploratoire, l'utilisateur aura à sa disposition une liste de fonctions (**Annexe 1**) qui combinés avec les connaissances des domaines vont constituer des stéréotypes. Il choisira celui qui lui conviendra, et le système répondra par des résultats mémorisés en fonction du stéréotype choisi.

La recherche précise sera effectuée à l'aide de mots-clés ou du texte intégral fourni par l'utilisateur durant une saisie libre. Elle agit forcément sur un segment logique, sémantique ou un document entier [ABA05a]. Comme les noms l'indiquent, une thèse a une structure logique telle l'introduction, le développement, la conclusion, les chapitres, les paragraphes etc., ainsi qu'une structure sémantique telle l'état de l'art, le prototype, la méthodologie etc.

Une fois la requête envoyée, l'utilisateur attend des réponses dont la modalité de restitution dépendra de ses préférences. C'est pourquoi nous avons identifié un dernier groupe de connaissances (CRS) qui sont présentées dans le paragraphe suivant.

3.1.6 Connaissances de restitution (CRS)

La réponse aux questions de l'utilisateur : « sous quelle apparence je veux mes réponses? » et sur « quel matériel? » constitue les éléments de définition des connaissances CRS. Nous allons seulement relever des éléments élémentaires nécessaires à notre étude. Donc, ces connaissances vont contenir les préférences documentaires, les adaptations nécessaires pour les handicapés ou pour les besoins spécifiques de l'utilisateur, ainsi que les périphériques de restitution (**Figure 7**).

Un certain nombre d'éléments ont été représentés dans cette typologie, toutefois nous nous intéressons seulement aux préférences documentaires de restitution sur un ordinateur fixe de bureau. Un travail de recherche, au sein de l'équipe par sur cette problématique que nous ne que partiellement dans notre sujet. De plus, étant donné qu'il s'agit purement de renseignements concernant l'affichage des résultats, ces éléments ne vont pas être pris en compte pour la constitution du cas. Ils seront utilisés juste avant de délivrer les résultats comme indiqué dans la Figure 9, étape 9.

A partir de cette typologie sur les connaissances de l'utilisateur, nous allons nous intéresser aux aspects formalisation, construction et évolution du profil utilisateur.

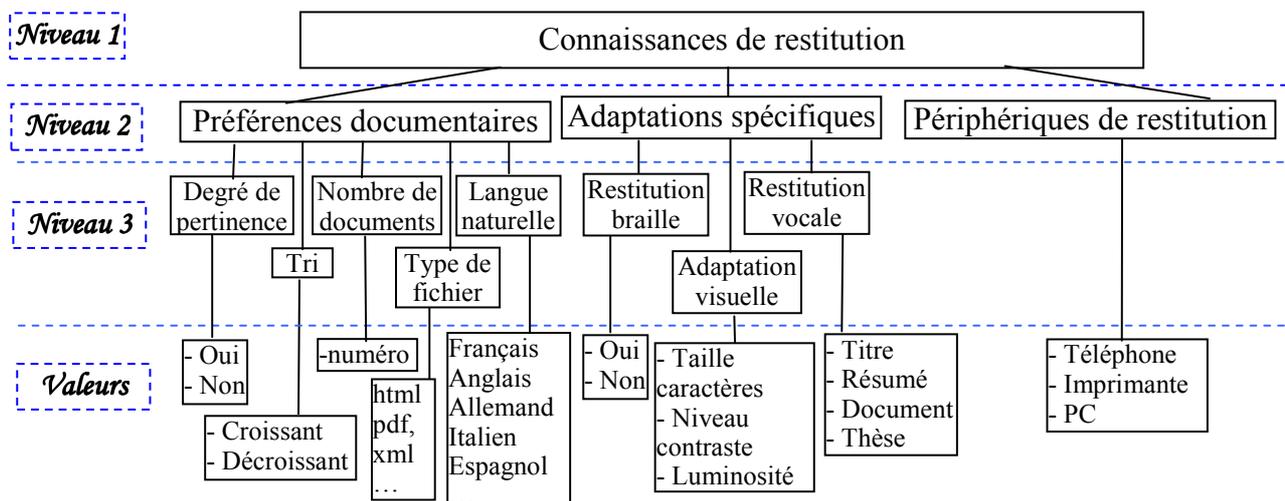


Figure 7. Modèle des connaissances de restitution

3.2. Formalisation et construction du modèle de l'utilisateur ainsi qu'évolution d'un profil

3.2.1. Description du cas

Le modèle de l'utilisateur sera formalisé par un cas. Nous avons indiqué (§ 2.4.4, modélisation du cas), que le cas sera représenté par une liste d'attributs – valeurs. Nous n'allons pas lister tous les attributs-valeurs associés à toutes les connaissances citées dans les paragraphes précédents. Certaines seront nécessairement renseignées par l'utilisateur, d'autres pourront être déduites en fonction des caractéristiques des stéréotypes. Par exemple, un enseignant a des droits de lecture sur toutes les thèses dont il est rapporteur, même avant la diffusion, alors qu'un utilisateur du système de RI, n'a que les droits de lecture sur les thèses après diffusion. C'est pourquoi, nous avons réfléchi attentivement aux connaissances que nous allons retenir pour définir un cas. Le but est d'avoir une liste, qui dans la mesure du possible, représente d'une manière significative les caractéristiques et comportements de la globalité de la population des utilisateurs. En choisissant ces éléments, nous nous sommes posé la question « quel élément doit-il obligatoirement être conservé ? ». L'intérêt porte sur deux axes : (i) quelles connaissances vont être réutilisées par l'utilisateur courant, et (ii) quelles connaissances sont susceptibles d'être réutilisées par d'autres utilisateurs.

3.2.2. Construction du cas

Nous avons vu (§2.4.4, Description du cas) qu'un cas peut décrire une situation de diagnostic contenant en général la description des symptômes, du défaut et sa cause, ainsi que la description de la stratégie de réparation. Nous nous conformons à ce cycle qui vise à faciliter l'écriture des requêtes de l'utilisateur en les enrichissant avec des données invariantes, communes à toutes les requêtes, mais spécifiques à un utilisateur, ou un groupe d'utilisateurs.

Pour notre étude, un cas sera construit par la situation réelle lors d'une session de recherche, le diagnostic que le système va faire à partir de cette situation et la réponse que le système va apporter.

La situation de recherche d'information concerne les connaissances réelles de l'utilisateur sur le système (CU), et la requête émise par l'utilisateur (RU). Une fois que le système connaîtra tous ces éléments, il va chercher à améliorer la requête de l'utilisateur en fonction de la précision de la requête émise. La requête améliorée (RA) va constituer le diagnostic. Ensuite, le système va effectuer une recherche d'information en fonction de la requête améliorée et va rendre une réponse à l'utilisateur, celui-ci aura alors la possibilité d'en évaluer la pertinence par une note de 1 à 10. La note de l'utilisateur sera stockée parmi les descripteurs du diagnostic. La réponse, qui sert de solution du cas sera constituée des documents (D) retournés par un système (Tableau 2).

	Grpe	Sous - Groupe	Attribut	Valeur groupe	Poids	Type de valeur
Situation réelle	CU	Identification(CG) Domaine (CD) Manipulation(CS)	Identifiant	1	2	Valeur précise
			Handicap			1
			Fonction		3	Valeur précise
			Domaine principal		4	Valeur précise
			Sous-domaine		4	Valeur précise
			Spécialité		3	Valeur précise
			Domaines Connexes		3	Valeur précise
			Fréquence utilisation		2	Valeur précise
	RU	(CRE)	Type de document	2	2	Règles de similarité
Niveau de spécialisation			2			Règles de similarité
Exploration			3			Similarité d'arbre
Rech. thématique			3			Similarité d'arbre
			3			Similarité d'arbre
Rech. précise			3			Valeur précise
			1			Valeur précise
			1			Règles de similarité
			3			Similarité d'arbre
Diagnose	RA	Requête améliorée	Type de document			
			Niveau de spécialisation			
			Thème			
			Sujet			
			Segment sémantique			
			Mots-clefs			
			Note utilisateur			
Solution	D		Titres de document			
			Auteur du document			
			Contenu du document			

Tableau 3. Structure d'un cas

Donc, la base sera constituée des cas représenté par

Cas Source = (CU, RU, RA, D) (1)

En se basant sur la typologie des connaissances telles que décrites plus haut le CU et RU se formalisent par

CU = (CG, CD, CS) et RU = CRE (2)

Conformément à (1) et (2) le cas de la base sera constitué ainsi

Cas Source = (CG, CD, CS, CRE, RA, D) (3)

Un cas cible, cas nouveau arrivant dans la base, ne sera constitué que des connaissances de l'utilisateur Cas Cible = (CU, RU) ou Cas Cible = (CG, CD, CS, CRE).

Avant que le système cherche la solution dans la base documentaire, on aura un cas intermédiaire constitué des connaissances et du diagnostic formalisé par le

Cas intermédiaire = (CU, RU, RA) ou Cas intermédiaire = (CU, RU, RA).

Il apparaît dans le Tableau 3 que les attributs sont classés et pondérés au sein d'un même cas afin de faciliter le travail algorithmique du système durant le calcul des similarités.

Donc, un exemple de cas stocké dans notre base et formalisé par

Cas Source = (CG, CD, CS, CRE, RA, D) sera

CG = ("00001 ", "Néant ")

CD= ("Etudiant", "Informatique ", "Connaissances et raisonnement", "Donnée, Document, Connaissances", "Bibliothèques, Archivage")

CS = ("Fréquemment ")

CRE = ("Thèse", "Débutant ", "NON", "Néant ", "Néant ", "Etat de l'art", "Néant ", "Expression exacte", "Raisonnement à partir des cas")

RA = ("Thèse, Master Recherche", "Intermédiaire", "Informatique : Intelligence artificielle", "Raisonnement à partir des cas", " Etat de l'art, méthodologie, raisonnement", "Raisonnement à partir de l'expérience")

3.2.3. Exécution adaptative des requêtes tenant compte des profils

Dans notre système de recherche de l'information, fondé sur l'exploitation de bases de cas, nous avons introduit dans le cas, des descripteurs appartenant aux connaissances de recherche, qui sont utiles pour l'enrichissement de la requête. Ainsi, les descripteurs que nous avons retenus pour l'expansion et la réécriture de la requête sont : type de document, niveau de spécialisation, thème, sujet, segment sémantique, mots-clés. Dans notre exemple, la RA indique que les éléments cherchés peuvent se trouver dans une Thèse ou un Master Recherche, et correspondent à un niveau Intermédiaire. Thèse et Master sont des valeurs de la connaissance de type de document. Dans ce cas la requête a été améliorée en ajoutant une valeur. En ce qui concerne le niveau de spécialisation, le système a modifié la valeur de Débutant en Intermédiaire. Apparemment, l'utilisateur n'estime pas bien son niveau de connaissance de domaine. Donc, dans la RA (la partie diagnostic), nous trouvons les valeurs enrichies de la requête grâce à l'adaptation des cas similaires. Les attributs du cas sont des données invariantes, communes à toutes les requêtes de tous les utilisateurs. Par contre, les valeurs attribuées durant une session de recherche sont spécifiques à un utilisateur ou un groupe d'utilisateurs. C'est ainsi que nous procédons à la personnalisation de la RI. Maintenant le but est d'évaluer l'impact de la modification de la requête sur le résultat, en terme de pertinence, et d'évaluer la qualité de la réponse. Pour réaliser cette évaluation, nous devons mettre en place le processus de raisonnement à partir des cas.

3.2.4 Caractéristiques des stéréotypes

Durant notre étude, il est apparu évident qu'un certain nombre de groupes d'utilisateurs interviennent dans le système de DOC'INSA. Les caractéristiques qui définissent ces groupes sont le "statut" et la "fonction". En résumé, les groupes régis par le "statut" sont les constructeurs, les bibliothécaires, les gestionnaires de la base et les invités. Les groupes régis par la fonction qu'ils

exercer à l'INSA sont les étudiants et les professeurs. Mais ces deux groupes se subdivisent selon une autre caractéristique, leur "activité". Ainsi, parmi les professeurs on trouve les directeurs de thèses, les rapporteurs et les membres de jury, et parmi les étudiants, nous trouvons ceux du premier cycle, ceux du cycle d'ingénieur (4-5IF, master pro, cnam), les pré-doctorants (master recherche), et les doctorants. Pour tous ces groupes on peut établir des plans d'actions (selon la proposition de KOBASA) différents à l'intérieur d'un même module, en se basant sur la caractéristique de "manipulation" de système. Par exemple, durant la mise en place d'une thèse, un doctorant la rédige, un directeur la corrige, une bibliothécaire l'archive et un gestionnaire du système gère le document dans la base.

Or, dans notre session de recherche d'information, les utilisateurs de tous ces groupes font la même utilisation du système, et donc, la manipulation n'est plus une caractéristique valable. C'est pourquoi, il faut prendre en compte des caractéristiques transverses à ces groupes pour définir une nouvelle manière de grouper. Ces types d'éléments sont les connaissances du domaine sur lequel les utilisateurs font la RI. Donc, les caractéristiques de nos nouveaux stéréotypes (nous avons déjà expliqué le lien entre groupe et stéréotype dans la partie portant sur l'approche sociologique) seront le "domaine", le "sous-domaine" la "spécialité" et les "champs connexes". Remarquez, qu'étrangement, ce sont des descripteurs de cas de forte importance.

Mais il faut également trouver une caractéristique qui fait le lien entre la première catégorisation et la deuxième. Pour nous, ce sera le "niveau de spécialisation". En effet un étudiant pré-doctorant, aura un niveau de spécialisation moins élevé dans un sujet précis, qu'un étudiant doctorant de troisième année ou encore un professeur travaillant sur ce même sujet. Ainsi, pour un utilisateur qui procède à une recherche exploratoire, par définition très imprécise en renseignant seulement le profil "pré-doctorant", le domaine informatique, et sous-domaine "connaissances et raisonnement", on va chercher tous les cas où le niveau de spécialisation est "intermédiaire" ou "débutant" dans une échelle de : « débutant », « intermédiaire », « avancé », « expérimenté », spécialiste. Le problème est de savoir, quels sont les sujets correspondant aux niveaux de spécialités choisis. C'est là que prend son importance le raisonnement à partir de cas. Dans un cas, nous avons décrits d'autres descripteurs que les caractéristiques des stéréotypes, comme par exemple la requête précise d'un utilisateur, ou encore, la requête améliorée. Il est à remarquer, que les requêtes sont les actions effectuées par les utilisateurs lors de leurs recherches. Ainsi, en stockant les cas, nous avons également stocké des plans d'actions pour les actions des stéréotypes, en sélectionnant tous les cas utilisés avec ce niveau de spécialisation.

Ainsi, le profil aura évolué d'un modèle d'utilisateur propre au système, à un modèle d'utilisateur propre au domaine de la recherche.

3.2.5. Processus du raisonnement

Comme nous l'avons déjà indiqué, le raisonnement à partir des cas comprend 4 étapes : retrouver, adapter, réviser et mémoriser.

Pour retrouver un cas, il faut mettre en place des mesures de similarité afin de procéder à des calculs statistiques pour le classement du nouveau cas parmi ceux de la base et la sélection du cas à adapter. C'est pourquoi, à droite du tableau 3, il y a 3 colonnes : la valeur d'un groupe, le poids des descripteurs du cas et les types de comparaisons à réaliser pour chaque descripteur. En ce qui concerne le poids, nous avons utilisé une échelle de 1 à 4 pour l'évaluation et nous avons jugé que les descripteurs les plus importants sont ceux qui caractérisent le profil de l'utilisateur que ce soit au niveau des connaissances du domaine ou des connaissances de la recherche.

En ce qui concerne le type de calcul de similarité, là où il y a « valeur exacte », la comparaison doit donner exactement la même chaîne de caractères et donc le seuil de similarité est de 100% (ou 10 pris d'une échelle de 1 à 10).

En ce qui concerne les règles de similarités, il s'agit de règles prédéfinies manuellement car les concepts manipulés ne sont pas nombreux. Par exemple, dans la comparaison des types de documents, il n'y a pas beaucoup de concepts à manipuler, nous avons donc défini des valeurs de similarités pris dans une échelle de 1 à 10. Donc, nous avons jugé que de part leur structure et la manière dont les contenus sont développés,

$\text{sim}(\text{"Doctorat"}, \text{"Mémoire CNAM"}) = 7$
 $\text{sim}(\text{"Doctorat"}, \text{"Master recherche"}) = 9$
 $\text{sim}(\text{"Master Recherche"}, \text{"Mémoire CNAM"}) = 8$
 $\text{sim}(\text{"Doctorat"}, \text{"Cours"}) = 5$

Certes pour commencer notre étude, nous avons pris des valeurs arbitraires qui, selon notre niveau de connaissance obéissent au bon sens, mais la construction d'une ontologie de ce domaine est un des points à améliorer pour l'avenir.

En ce qui concerne les similarités d'arbre, il s'agit de calculer une distance entre concepts qui sont déjà catégorisés dans une base de concepts ou une ontologie.

Donc, la similarité entre deux cas sera exprimée par la formule

$$\text{Sim}(\text{cas } s, \text{cas } c) = \left(\frac{\sum p_i * \text{sim}(d_{is}, d_{ic})}{\sum p_i} \right)$$

Où, i est le rang de chaque descripteur dans le cas, pi le poids de ces descripteurs, et d représente la distance.

Donc, on calcule la similarité à l'intérieur de chaque groupe et ensuite on attribue la valeur du groupe.

L'algorithme du calcul du plus proche voisin est le suivant:

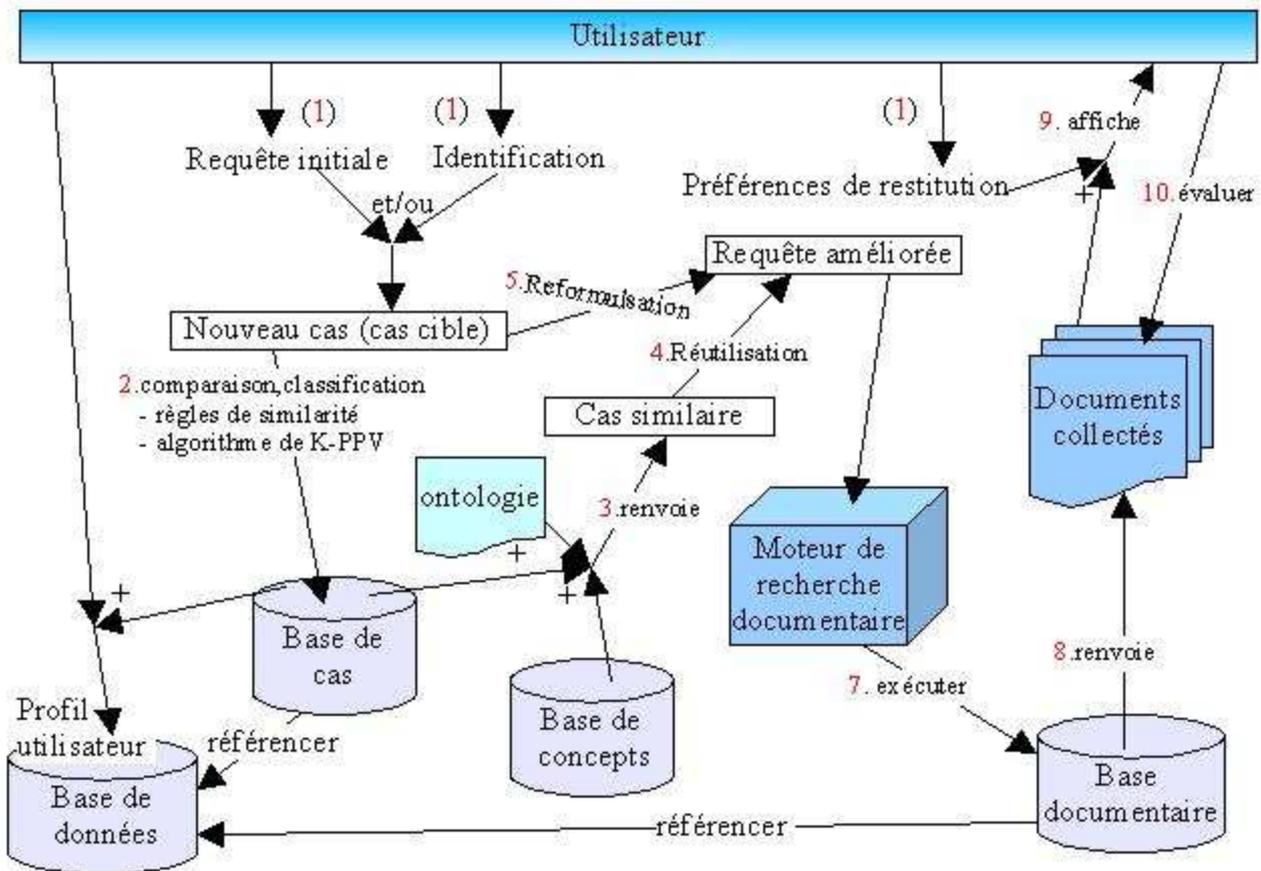


Figure 9. Structure du système

3.2.6. Architecture du système

En fonction des actions enchaînées durant leur processus de raisonnement à partir des cas, nous avons construit la structure architecturale de notre système.

Pour commencer, le système contient 4 bases : la base de données, la base des cas, la base des concepts et la base documentaire. Les cas et les documents sont référencés également dans la base de données.

Pour commencer (étape 1) l'utilisateur saisit ses renseignements dans le système grâce aux écrans de la Figure 10. S'il est un utilisateur reconnu de l'INSA, il s'identifie tout d'abord, sinon, il passe directement à la saisie de sa requête où il renseigne les connaissances de la recherche et ses préférences de livraison. Ainsi la liste des attributs du nouveau cas cible aura des valeurs renseignées. Grâce aux mesures de similarité et à l'algorithme du K-PPV, le système va classer le cas cible. Ensuite grâce à l'algorithme du K-PPV et à l'utilisation d'une ontologie de domaine et à une base de concepts, le système va choisir le meilleur cas de la classe à réutiliser. La requête apparaissant dans le diagnostic du cas similaire sera adaptée au cas cible, et ainsi le système aura reformulé la requête de l'utilisateur avec laquelle il va procéder à une recherche d'information dans la base documentaire. Cette dernière étape, est similaire aux problèmes classiques de fouilles de données dans des espaces documentaires, mais ceci ne constitue pas l'objet de notre sujet. C'est pourquoi, nous prenons un moteur de recherche disponible. Si aucun cas similaire n'a pas pu être trouvé, le système va malgré tout procéder à une recherche d'information dans la base documentaire, mais sans améliorer la requête en fonction du profil. Ensuite, le système va retourner les documents sélectionnés qui seront affichés selon les préférences définies par l'utilisateur. Enfin, il aura la possibilité d'évaluer la pertinence de ces réponses sur une échelle de 1 à 10. Enfin, si un cas est utilisé au-delà d'un seuil S fixé après analyse du système, par l'utilisateur, ou si tous les descripteurs de caractéristiques des stéréotypes ont été renseignés, le cas sera stocké parmi les données de l'utilisateur en tant que profil de l'utilisateur. Pour des questions de maintenance, nous allons nous limiter au stockage de 5 cas différents dans un premier temps.

Chapitre 4. Evaluation

Actuellement, CITHER offre les possibilités de recherche suivantes :

- Combinaison de recherche par année de soutenance et nom d'auteur
- Combinaison de recherche par auteur, mot du titre, cote, et année d'édition
- Par mots-clés dans le texte intégral, définies au mieux au niveau des chapitres

Nous allons nous intéresser plus particulièrement à la recherche par mots-clés, car cette option nous permet de mieux mettre en évidence les avantages que propose notre système.

Si l'utilisateur effectue une recherche par le mot-clé « réseau sémantique », il va avoir le résultat de la Figure 10. Donc, le système a répondu avec tous les documents qui traitent de réseau sémantique. Or, si on observe plus précisément les résultats, on constate que la première thèse traite du domaine connexe de la linguistique, la deuxième thèse dans la liste, la troisième, cinquième et sixième concernent le domaine médical (d'ailleurs les trois lignes correspondent à la même thèse sans avoir plus de précisions). La quatrième traite de notions techniques d'informatique fondamentale, la septième concerne le monde multimédia, et la dernière prend comme champ d'application les documents archéologiques.

Tout ceci signifie, qu'un utilisateur qui souhaite traiter des documents archéologiques, et qui n'aura pas su préciser sa requête, aura obtenu toutes ces réponses au lieu de l'unique qui intéresse vraiment. D'ailleurs, les documents de notre exemple précisent clairement dans le titre le domaine d'application, mais s'il n'est pas précisé, ce qui est fréquent, l'utilisateur doit lire le contenu de tous les documents pour savoir de quoi, il s'agit.

A l'inverse, avec notre système, l'utilisateur va pouvoir trouver la réponse pertinente Figure 11. Ceci est expliqué par deux cas.

- a- Un utilisateur non identifié par le système recherche de l'information en précisant ses connaissances de domaine
- b- Un utilisateur identifié (travaillant sur les documents archéologiques) saisie le mot-clé « réseau sémantique »

Dans le premier cas de figure, l'utilisateur va être assisté pour renseigner son domaine principal, qui pour notre exemple sera l'Informatique, et son domaine d'application qui sera l'Archéologie. Ensuite, le système va chercher le stéréotype caractérisé par ces deux valeurs et va pouvoir renseigner quel type de documents doit être retourné à un tel utilisateur.

La recherche avancée (avec profil utilisateur) de ce cas est présentée en Annexe 2.

The screenshot shows the CITHER search interface. At the top, there are logos for INSA Lyon and CITHER, along with the text 'Bibliothèques de L'INSA de LYON'. The main content area is titled 'En texte intégral' and contains a search box with the text 'réseau sémantique' and a 'Recherche' button. Below the search box, there is a list of search results, each with a title, author, and file size. The results are as follows:

Recherche: r?seau s?mantique*
9 document(s) trouvé(s)
Une méthode d'indexation sémantique adaptée aux corpus multilingues
Auteur: ROUSSEY Catherine
Taille du fichier: 1462994 octets
Utilisation des Topic Maps pour l'interrogation et la génération de documents virtuels : Application au domaine médical
Auteur: OUZIRI Mourad
Taille du fichier: 111665 octets
Utilisation des Topic Maps pour l'interrogation et la génération de documents virtuels : Application au domaine médical
Auteur: OUZIRI Mourad
Taille du fichier: 508224 octets
SAGED-XML : serveur actif pour la gestion de la cohérence de documents
Auteur: ALVAREZ ESCOBEDO Abraham
Taille du fichier: 347320 octets
Utilisation des Topic Maps pour l'interrogation et la génération de documents virtuels : Application au domaine médical
Auteur: OUZIRI Mourad
Taille du fichier: 386878 octets
Utilisation des Topic Maps pour l'interrogation et la génération de documents virtuels : Application au domaine médical
Auteur: OUZIRI Mourad
Taille du fichier: 26927 octets
Gestion des connaissances dans une base de documents multimédias
Auteur: EGYED-ZSIGMOND Elod
Taille du fichier: 3216966 octets
Consultation assistée par ordinateur de la documentation en Sciences Humaines : Considérations épistémologiques, solutions opératoires et applications à l'archéologie
Auteur: BENEL Aurélien
Taille du fichier: 3584179 octets
Modélisation multiparadigme de textes réglementaires
Auteur: Bertrand Chabbat, Email : bchabbat@usa.net
Taille du fichier: 2201869 octets

At the bottom of the browser window, the status bar shows 'Terminé'.

Figure 10. Recherche de l'information en texte intégral avec le système actuel

Dans le deuxième cas de figure, étant donné que l'utilisateur s'identifie dans le système, il est reconnu par le système qui a déjà stocké un stéréotype associé à son profil. Il s'agira du stéréotype caractérisé par la valeur du domaine principal « Informatique » et par la valeur du domaine connexe « archéologie ». Rappelons ici, que le stéréotype sera stocké parmi les données de l'utilisateur, dans la base de données, sous la forme d'un cas.

De cette manière, avec une quantité minimum d'informations, notre système pourra renvoyer des réponses pertinentes à ses utilisateurs.

Chapitre 5 : Conclusion et perspectives

La prise en compte des besoins, des intentions, et des spécificités cognitives, culturelles ou autres qui caractérisent le profil de l'utilisateur constituent un élément déterminant pour améliorer la pertinence des réponses lors d'une session de Recherche de l'information dans de grandes bases de documents. Dans ce mémoire, nous avons présenté une des pistes que nous explorons pour proposer un accès pertinent au contenu de ces documents, qui est la définition et l'exploitation du profil utilisateur en vue de satisfaire au mieux la demande de l'utilisateur dans ce contexte de RI.

Après avoir réalisé un état de l'art permettant de dégager les connaissances constitutives d'un profil utilisateur ainsi que leurs modes d'acquisition, de structuration et d'évolution, nous avons proposé un modèle fondé sur les connaissances de l'utilisateur adapté au domaine de la recherche d'information. En raison des avantages offerts et de son adéquation à nos besoins spécifiques, nous avons implémenté la technologie du Raisonnement à Partir de Cas pour structurer, acquérir et faire évoluer le profil de l'utilisateur. De plus, nous avons dégagé un certain nombre de stéréotypes d'utilisateurs qui nous permettent de démarrer le processus du RaPC et de bien cerner les caractéristiques et les comportements des groupes d'utilisateurs utilisant notre système.

Ceci constitue une première étape qui a conduit à la définition d'un modèle et d'un prototype qui doit être plus largement éprouvé.

Il reste encore des points d'amélioration qui concernent:

1. La réalisation de tests à une plus grande échelle afin que les résultats soient plus représentatifs.
2. Le développement complet de la caractérisation des stéréotypes.
3. La construction de l'ontologie du domaine.

En ce qui concerne le premier point, le corpus sur lequel les tests ont été réalisés, est constitué d'une vingtaine de thèses numériques du domaine de l'informatique. Le but est de continuer l'évaluation, à l'échelle de tous les domaines de CITHER, afin de mieux se rendre compte des avantages qu'apporte la personnalisation en termes de gain de pertinence de l'information recherchée. Ces tests permettraient également d'enrichir la base de cas, ce qui revient à enrichir les connaissances du système sur le domaine. Ceci nous permet d'aborder le deuxième point d'amélioration.

Durant notre étude nous avons tâché de déduire un certain nombre de stéréotypes en se appuyant sur notre expérience personnelle d'utilisation de l'outil de DOC'INSA. Bien évidemment, que cela ne vaut pas l'expérience acquise par tous les utilisateurs de CITHER. Il est certain qu'à l'heure actuelle il nous est impossible de mettre en relation toutes ces connaissances afin d'établir des plans d'actions pour chaque stéréotype. Or, notre système pourra gagner toute cette expertise durant des tests sur un corpus plus important et avec un grand nombre d'utilisateurs. A partir de cette expertise, il sera certainement possible d'inférer d'autres caractéristiques et d'autres plans d'action en mettant ainsi en évidence de nouveaux stéréotypes plus précis.

Afin d'initialiser notre système, nous avons convenu de se conformer à l'organisation de l'enseignement de l'INSA dans le choix des domaines, sous-domaines, et spécialités. Certes, cela satisfait nos premiers besoins d'évaluation, mais nous sommes bien conscients qu'un tel système nécessite une ontologie élaborée du domaine pour fonctionner en mode réel et à grande échelle.

Bibliographie

- [**ABA04**] Abascal R., Rumpler B., Pinon J.M., (2004) Information Retrieval in Digital Theses Based on Natural Language Processing Tools, J.L. Vicedo et al. (Eds): España for Natural Language Processing (EsTAL'04), LNAI 3230, pp. 172-182, Springer-Verlag Berlin Heidelberg, October 2004, Alicante, Spain.
- [**ABA05a**] Abascal R., Rumpler B., Berisha-Bohé S., Pinon J-M. "A Semantic Structure for Digital Theses Collection Based on Domain Annotations", 9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2005). Austria, September, 2005.
- [**ABA05b**] Abascal R., Rumpler B., Berisha-Bohé S. "Proposition d'une nouvelle structure de document pour améliorer la recherche d'information", Proceedings of the CORIA'05 (CONFérence en Recherche d'Informations et Applications), ISBN: 2-9523810-0-3, IMAG, pp. 389-404, 2005.
- [**BEN93**] Benhamou F., Colmerauer A., Constraint Logic Programming: Selected Research, MIT Press, 1993, ISBN 0-262-02353-9
- [**BOH05**] Berisha-Bohé S., Abascal R., Rumpler B., A semantic structure to improve information retrieval using XML,
- [**CRA02**] Crampes Michel, Auto-Composition Active et émergence du sens dans l'interaction Homme-Machine sous contrainte, HDR, 2003
- [**CRIS02**] Cristea Alexandra I, User Modelling meets the web, Modern Information Systems seminar, 26th February 2002, www.wis.win.tue.nl/~acristea/presentations/UMpres.ppt
- [**FINK00**] Fink.J. And Kobsas. A., A Review and Analysis of Commercial User Modeling Servers for Personalization on the World Wide Web, User Modeling and User-Adapted Interaction (10), Kluwer Academic Publishers, pp. 209-249, 2000.
- [**FINK02**] Fink.J. And Kobsas. A., User Modeling for Personalized City Tours, Artificial Intelligence Review 18, 2002, pp.33-74, Kluwer Academic Publishers
- [**Gaus03**] Gaussier, E. Stefanini, MH. Assistance intelligente à la recherche d'informations, Hermes Science, ISBN 2-7462-0726-5, 2003
- [**GIU92**] Giunchiglia F., Toby Walsh: A Theory of Abstraction, Artificial Intelligence 57, 1992
- [**KAY01**] Kay, J, R J Kummerfeld and P Lauder, Foundations for personalised documents: a scrutable user Model Server, *Proceedings of ADCS'2001, Australian Document Computing Symposium*, pp. 43-50, 2001
- [**KAY99**] Kay, J., Ontologies for reusable and scrutable student model, position paper, In Proceedings of AIED99 Workshop on Ontologies for Intelligent Educational Systems.
- [**KOB01**] Kobsa. A., Generic User Modelling Systems, User Modeling and User-Adapted Interaction (11), Kluwer Academic Publishers, pp. 49-63, 2001
- [**KOB03**] Kobsa. A. and Fink.J, Performance Evaluation of User Modeling Servers Under Real- World Workload Conditions, 9th international Conference on User Modelling, Johnstown, PA, 2003
- [**KOL93**] Kolodner, J., Case-based Reasoning, Morgan Kaufmann Publishers, 1993
- [**LAM02**] Lamontagne. Luc. And Lapalme. G., Raisonnement à base de cas textuels-état de l'art et perspectives, RSTI série RIA, Volume 16-n°3, 2002, pages 339 à 366
- [**MID03**] Middleton, SE., Capturing knowledge of user preferences with recommender systems, Thesis, University of Southampton, 2003.
- [**NOS84**] NOSOFSKY J. 1984. Choice, Similarity, and the context theory of classification.
- [**RICH79**] Rich E., User Modeling via Stereotypes, Cognitive Science, 3, pp. 329-354 (1979)
- [**RICH83**] Rich, E., Users are individuals:- individualizing user model, International Journal Man Machine Studies (1983) 18, pp. 199-214
- [**RICH89**] Rich, E., Stereotypes and user modeling. In: A. Kobsa, and W. Wahlster (eds.), User Models in Dialog Systems. Springer, Berlin, Heidelberg, pp. 35-51, (1989) .
- [**RUS03**] [Russell S., Norvig P., Artificial Intelligence, A Modern Approach: Second Edition](#), Prentice-Hall, 2003
- [**SCH89**] SCHANK, C. Roger, RIESBECK, K. Christopher, Inside Case-Based Reasoning, LEA Publishers, Hillsdale, New Jersey, 1989
- [**SOLL03**] Sollenborn. M., Licentiate Thesis Proposal Clustering and Case-Based Reasoning for User Stereotypes (2003).
- [**ZUK01**] Zukerman, I. and Albrecht, D. (2001), [Predictive Statistical Models for User Modeling](#), User Modeling and User-Adapted Interaction, 11(1-2), 5-18. Invited paper <http://portal.acm.org/citation.cfm?id=598284.598344>
- Journal of Experimental Psychology: Learning, Memory, and Cognition*, Volume 10. pp. 104-114.

Annexes

Annexe 1 : Listes

Types de documents

1. Catalogues
2. Bibliographie sciences
3. Bibliographie humanités
4. Dictionnaires
5. Encyclopédies
6. Fiches techniques
7. Documents officiels
8. Annuaires
9. Presse actualité
10. Congrès
11. Thèses
12. Revues électroniques
13. Signets Internet
14. Brevets, Normes
15. Cours
16. Catalogues produits

Domaines connexes (dans le prototype actuel)

1. Archéologie
2. Architecture
3. Médecine
4. Bibliothèques
5. Linguistique
6. Multimédia

Fonctions

1. Etudiant 1er cycle
2. Pré-doctorant
3. Cycle Ingénieur
4. Ingénieur CNAM
5. Doctorant
6. Directeur de thèse
7. Enseignant
8. Rapporteur
9. Bibliothécaire
10. Administrateur Système
11. Utilisateur RII

¹ Utilisateur qui utilise le système qu'en session de recherche

Domaines (prises en compte dans le prototype actuel)

1. Biosciences
2. Génie Civil et Urbanisme
3. Génie Electrique
4. Génie Energetique et environnement
5. Génie Industriel
6. Génie Mécanique Conception
7. Génie Mécanique Développement
8. Informatique
9. Science et génie des matériaux
10. Télécommunications, Services et Usages

Spécialités (prises en compte dans le prototype actuel)

1. Données, documents et connaissances
2. Images et vidéos : segmentation et extraction d'information
3. Modélisation et réalité augmentée
4. Systèmes d'information communicants

Sous domaines de l'informatique (prises en compte dans le prototype actuel)

1. Connaissances et Raisonnement
2. Aide à la Décision pour l'Entreprise
3. Extraction de Connaissances à partir des Données
4. Informatique graphique et images
5. Systèmes d'information
6. Réseaux, Télécommunications et Services
7. Informatique Fondamentale

Annexe 2 : Ecrans du prototype

Exemples de quelques écrans correspondants à a recherche d'information avancée géant le profil utilisateur. Nouvelles fonctionnalités rajoutées à CITHER.

Définir votre profil

Vous êtes un:

Identification

login:	<input type="text"/>
mot de pass:	<input type="text"/>
<i>existe seulement si c'est un nouvel utilisateur:</i>	
confirmer mot de pass:	<input type="text"/>

Sélection des domaines

Votre domaine principale:

Sous-domaine:

Votre spécialité:

Domaines connexes

deviennent actives après la sélection du domaine principal

Votre 1er domaine d'application:

Votre 2eme domaine d'application:

Votre 3eme domaine d'application:

Vous souhaitez accéder à une session de:

- Production de thèse Recherche d'information
 Archivage Gestion de système

Rechercher

Ecran de recherche (préférences)

Type de document: Élément documentaire:

Sujet:

Contenant les mots des:

Date de la thèse:

En fonction de renseignements de la page précédente:

Langue de la thèse (en écriture):

Préférences documentaires de restitution

Langue:

Format:

Périphérique:

Nombre de documents par page: Ordre:

Tri par:

Type de restitution:

Adaptation visuelle

Taille caractères:

Contraste:

Luminosité:

Search results: r?seau s?mantique* - Mozilla Firefox
Fichier Edition Affichage Aller à Marque-pages Outils ?
http://pcdoc-41.insa-lyon.fr/asp/dtsearch.asp?cmd=search&SearchForm=%25%25SearchForm%25%25&request=r%E... OK
Démarrage Dernières nouvelles ...



INSA-Lyon : Doc'INSA
Recherche en texte intégral

Recherche: r?seau s?mantique*	Degré de pertinence	Votre appréciation
1 document(s) trouvé(s) Modélisation multiparadigme de textes réglementaires Auteur: Bertrand Chabbat. Email : bchabbat@usa.net Taille du fichier: 2201869 octets	90%	<input type="checkbox"/>

Terminé